Theory of Probability revisited:
A reassessment of a Bayesian classic

Christian P. Robert

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

March 8, 2006

Outline

First chapter: Fundamental notions

Who's 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.

Jeffreys and Science

Jeffreys married another mathematician and physicist, Bertha Swirles (1903-1999) and together they wrote *Methods of Mathematical Physics*.

Jeffreys is the founder of modern British geophysics. Many of his contributions are summarised in his book *The Earth*. One of his main discoveries is that the core of the Earth is liquid.

Jeffreys and Statistics

Jeffreys wrote more than 400 papers, mostly on his own, on subjects ranging across celestial mechanics, fluid dynamics, meteorology, geophysics and probability.

> *H. Jeffreys and B. Swirles (eds.) (1971–77)* Collected Papers of Sir Harold Jeffreys on Geophysics and other Sciences in six volumes, *London, Gordon & Breach.*

The statistics papers are in volume 6, *Mathematics, Probability & Miscellaneous Other Sciences*. The coverage is not comprehensive for Jeffreys omitted papers that had been superseded by his books *Scientific Inference* and *Theory of Probability*.

Jeffreys and Inference

Jeffreys first used probability to deal with problems in the Philosophy of Science.
K. Pearson's *Grammar of Science* made a great impression on him, with its emphasis on the probabilistic basic of scientific **inference**.
Jeffreys treated probability as a degree of reasonable belief, an epistemic conception common to several Cambridge philosophers, including J.M. Keynes. He used probability to explicate **induction** and investigate the reasonableness of scientific theories.

For appraising scientific theories, Venn's probability as a limiting **frequency** was useless but Jeffreys considered it mathematically unsound as well.
Around 1930 Jeffreys began devising methods for analysing geophysical data based on epistemic probability. He was extending the methods used by physical scientists and did not know much about, or greatly esteem, the efforts of statisticians.

Jeffreys and 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 and Jeffreys first took serious notice of each another in 1933. About all they knew of each other's work was that it was founded on a **flawed notion of probability.**

Theory of Probability revisited
Fundamental notions
Sir Harold Jeffreys

The Fisher–Jeffreys controversy

The *Biometrika* papers

Jeffreys (1933a) criticised Fisher (1932) and Fisher (1933) criticised Jeffreys (1932) with a rejoinder by Jeffreys (1933b). *Biometrika* called a halt to the controversy by getting the parties to coordinate their last words, in Fisher (1934) and Jeffreys (1934).

Jeffreys' Theory of Probability

While Jeffreys conceded nothing to Fisher, the encounter affected the course of his work. He reacted to the dose of Statistics Fisher administered by reconstructing Fisher's subject on his own foundations.
Theory of Probability (1939) was the outcome, as a theory of inductive inference founded on the principle of inverse probability, **not** a branch of pure mathematics, **not** a description of natural phenomena as with Kolmogorov and von Mises.

Theory of Probability revisited
Fundamental notions
Theory of Probability

Theory of Probability begins with probability, refining the treatment in *Scientific Inference*, and proceeds to cover a range of applications comparable to that in Fisher's book.
Jeffreys was very impressed by the solutions Fisher had found for many statistical problems—**the trouble was that they had no real foundations!** He also tried to place Fisher's creations like **sufficiency** in his own system.

Theory of Probability?

First chapter **Fundamental Notions** sets goals for a theory of **induction** rather than the mathematical bases of **probability**

- Objection to Kolmogorov's axiomatic definition

  The above principles (...) rule out any definition of probability that attempts to define probability in terms of infinite sets of possible observations (I, §1.1, 8).

- No measure theoretic basis, e.g.

  If the law concerns a measure capable of any value in a continuous set we could reduce to a finite or an enumerable set (I, §1.62).

Theory of Probability revisited
Fundamental notions
Theory of Probability

- Logic based axioms (I, §1.2)
- Tautological proof of Bayes' Theorem

$$P(q_r|pH) \propto P(q_r|H)P(p|q_rH)$$

  where H is the information already available, and p some information (I, §1.22). This is the principle of inverse probability, *given by Bayes in 1763.*

- Introduction of decision theoretic notions like Laplace's *moral expectation* and utility functions
- Insistence on modelling and prior construction

Theory of Probability revisited
Fundamental notions
Theory of Probability

Conditional probabilities

Probabilities of events defined as degrees of **belief** and conditional on past **(prior)** experience

Our fundamental idea will not be simply the probability of a proposition p but the probability of p on data q (I, §1.2).

Subjective flavour of probabilities due to different data, $P(p|q)$, with same classical definition, e.g.

$$P(RS|p) = P(R|p)P(S|Rp)$$

proved for uniform distributions on finite sets (equiprobable events)

Theory of Probability revisited
Fundamental notions
Theory of Probability

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's 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*.

(Modern) parametric model

Observations $x_1, \ldots, x_n$ generated from a probability distribution

$$x = (x_1, \ldots, x_n) \sim f(x|\theta), \qquad \theta = (\theta_1, \ldots, \theta_n)$$

Fisher's associated likelihood

$$\ell(\theta|x) = f(x|\theta)$$

[inverse density]

Bayesian perspective

Jeffreys' premises

- *Prior beliefs* on the parameters θ of a model modeled through a *probability* distribution π on Θ, called *prior distribution*
- *Inference* based on the distribution of θ conditional on x, $\pi(\theta|x)$, called *posterior distribution*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, d\theta} \,.$$

The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods (I, §.1.22).

Modern Bayesian representation

Definition (**Bayesian model**)

A Bayesian statistical model is made of a parametric statistical model,

$$(\mathcal{X}, f(x|\theta))\,,$$

and a prior distribution on the parameters,

$$(\Theta, \pi(\theta))\,.$$

Jeffreys' Justifications

- All probability statements are conditional
- Actualization of the information on θ by extracting the information on θ contained in the observation x

  The principle of inverse probability does correspond to ordinary processes of learning (I, §1.5)

- Allows incorporation of imperfect information in the decision process

  A probability number [sic!] can be regarded as a generalization of the assertion sign (I, §1.51).

- Unique mathematical way to condition upon the observations (conditional perspective) [Jeffreys?]

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?

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

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

i.e.

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

[Beta distribution]

Prior and posterior distributions

Given $f(x|\theta)$ and $\pi(\theta)$, several distributions of interest:

(a) the *joint distribution* of (θ, x),

$$\varphi(\theta,x) = f(x|\theta)\pi(\theta)\,;$$

(b) the *marginal distribution* of x,

$$m(x) = \int \varphi(\theta,x)\,d\theta = \int f(x|\theta)\pi(\theta)\,d\theta\,;$$

(c) the *posterior distribution* of θ,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} = \frac{f(x|\theta)\pi(\theta)}{m(x)}\,;$$

(d) the *predictive distribution* of y, when $y \sim g(y|\theta,x)$,

$$g(y|x) = \int g(y|\theta,x)\pi(\theta|x)d\theta\,.$$

Prior Selection

First chapter of **ToP** quite obscure about choice of π
Seems to advocate use of uniform priors:

- *If there is originally no ground to believe one of a set of alternatives rather than another, the prior probabilities are equal (I, §1.22).*
- *To take the prior probabilities different in the absence of observational reason would be an expression of sheer prejudice (I, §1.4).*

Prior Selection (2)

Still perceives a potential problem:

> *...possible to derive theorems by equating probabilities found in different ways (...) We must not expect too much in the nature of a general proof of consistency (I, §1.5).*

but evacuates the difficulty:

> *...the choice in practice, within the range permitted, makes very little difference to the results (I, §1.5).*

Posterior distribution

- Operates conditional upon the observations
- Incorporates the requirement of the Likelihood Principle

  > *...the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give the likelihood (II, §2.0).*

- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected

  > *...can be used as the prior probability in taking account of a further set of data, and the theory can therefore always take account of new information (I, §1.5).*

Additional themes in **ToP** Chapter 1

- General remarks on model choice and the pervasive **Occam's rasor rule**
- Bayes factor for testing purposes
- Utility theory that evaluates decisions
- Fairly obscure disgressions on Logic and Gödel's Theorem.

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.

Second chapter: Direct Probabilities

Contents

Description and justification of most standard distributions
- Hypergeometric, Binomial, Negative Binomial, Multinomial
- Poisson
- Normal, Pearson, χ^2, Student's t

Comments

Physicist's approach (approximations, intuition, series expansion)
Strange mix of Math (more measure theory than in Chapter I) and pseudo-common sense

> *The normal law of error cannot therefore be theoretically proved (II, §2.68).*

Use of χ^2 test in a frequentist sense!
Advocates normal distributions on the Fourier coefficients

Prior remarks on prior Distributions

The most critical and most criticized point of Bayesian analysis !
Because...

the prior distribution is the key to Bayesian inference

But...

In practice, it seldom occurs that the available prior information is precise enough to lead to an exact determination of the prior distribution

There is no such thing as *the* prior distribution!

Rather...

The prior is a tool summarizing available information as well as uncertainty related with this information,

And...

Ungrounded prior distributions produce unjustified posterior inference

Subjective priors

In situations with prior information, choice of prior mostly subjective.

Example (Capture probabilities)

Capture-recapture experiment on migrations between zones

Prior information on capture and survival probabilities, p_t and q_{it}

	Time	2	3	4	5	6
p_t	Mean	0.3	0.4	0.5	0.2	0.2
	95% cred. int.	[0.1,0.5]	[0.2,0.6]	[0.3,0.7]	[0.05,0.4]	[0.05,0.4]

	Site	A		B	
	Time	t=1,3,5	t=2,4	t=1,3,5	t=2,4
q_{it}	Mean	0.7	0.65	0.7	0.7
	95% cred. int.	[0.4,0.95]	[0.35,0.9]	[0.4,0.95]	[0.4,0.95]

Example (Capture probabilities (2))

Corresponding prior modeling

Time	2	3	4	5	6
Dist.	$\mathcal{B}e(6, 14)$	$\mathcal{B}e(8, 12)$	$\mathcal{B}e(12, 12)$	$\mathcal{B}e(3.5, 14)$	$\mathcal{B}e(3.5, 14)$

Site	A		B	
Time	t=1,3,5	t=2,4	t=1,3,5	t=2,4
Dist.	$\mathcal{B}e(6.0, 2.5)$	$\mathcal{B}e(6.5, 3.5)$	$\mathcal{B}e(6.0, 2.5)$	$\mathcal{B}e(6.0, 2.5)$

Strategies for prior determination

- Use a partition of Θ in sets (e.g., intervals), determine the probability of each set, and approach π by an *histogram*
- Select significant elements of Θ, evaluate their respective likelihoods and deduce a likelihood curve proportional to π
- Use the *marginal distribution* of x,

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)\,d\theta$$

- Empirical and *hierarchical* Bayes techniques

Theory of Probability revisited
Direct Probabilities
Subjective determination

- Select a **maximum entropy** prior when prior characteristics are known:

$$\mathbb{E}^\pi[g_k(\theta)] = \omega_k \qquad (k = 1, \ldots, K)$$

with solution, in the discrete case

$$\pi^*(\theta_i) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta_i)\right\}}{\sum_j \exp\left\{\sum_1^K \lambda_k g_k(\theta_j)\right\}},$$

and, in the continuous case,

$$\pi^*(\theta) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta)\right\}\pi_0(\theta)}{\int \exp\left\{\sum_1^K \lambda_k g_k(\eta)\right\}\pi_0(d\eta)},$$

the λ_k's being Lagrange multipliers and π_0 a reference measure [Caveat]

Theory of Probability revisited
Direct Probabilities
Subjective determination

- **Parametric approximations**
  Restrict choice of π to a *parameterised* density

$$\pi(\theta|\lambda)$$

and determine the corresponding (hyper-)parameters

$$\lambda$$

through the *moments* or *quantiles* of π

Theory of Probability revisited
Direct Probabilities
Subjective determination

Example

For the normal model $x \sim \mathcal{N}(\theta, 1)$, ranges of the posterior moments for fixed prior moments $\mu_1 = 0$ and μ_2.

μ_2	x	Minimum mean	Maximum mean	Maximum variance
3	0	-1.05	1.05	3.00
3	1	-0.70	1.69	3.63
3	2	-0.50	2.85	5.78
1.5	0	-0.59	0.59	1.50
1.5	1	-0.37	1.05	1.97
1.5	2	-0.27	2.08	3.80

[Goutis, 1990]

Conjugate priors

Specific parametric family with analytical properties

Definition

A family $\mathscr{F}$ of probability distributions on Θ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathscr{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathscr{F}$.

[Raiffa & Schlaifer, 1961]

Only of interest when $\mathscr{F}$ is *parameterised* : switching from prior to posterior distribution is reduced to an updating of the corresponding parameters.

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$
- Exchangeability motivations
- Device of virtual past observations
- Linearity of some estimators
- Tractability and simplicity
- First approximations to adequate priors, backed up by robustness analysis

Theory of Probability revisited
Direct Probabilities
Conjugate priors

Exponential families

Definition

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\}$$

is called an *exponential family of dimension* k. When $\Theta \subset \mathbb{R}^k$, $\mathscr{X} \subset \mathbb{R}^k$ and

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta \cdot x\},$$

the family is said to be *natural*.

Interesting analytical properties :

- Sufficient statistics (Pitman–Koopman Lemma)
- Common enough structure (normal, binomial, Poisson, Wishart, &tc...)
- Analycity ($\mathbb{E}_\theta[x] = \nabla\psi(\theta)$, ...)
- Allow for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta\cdot\mu - \lambda\psi(\theta)}$$

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu+\tau^2 x),\rho\sigma^2\tau^2)$ $\rho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+\nu,\beta+x)$
Binomial $\mathcal{B}(n,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$

Theory of Probability revisited
Direct Probabilities
Conjugate priors

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Negative Binomial $\mathcal{N}eg(m,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Theory of Probability revisited
Direct Probabilities
Conjugate priors

Linearity of the posterior mean

If

$$\theta\sim\pi_{\lambda,x_0}(\theta)\propto e^{\theta\cdot x_0-\lambda\psi(\theta)}$$

with $x_0\in\mathcal{X}$, then

$$\mathbb{E}^\pi[\nabla\psi(\theta)]=\frac{x_0}{\lambda}.$$

Therefore, if $x_1,\ldots,x_n$ are i.i.d. $f(x|\theta)$,

$$\mathbb{E}^\pi[\nabla\psi(\theta)|x_1,\ldots,x_n]=\frac{x_0+n\bar{x}}{\lambda+n}.$$

Theory of Probability revisited
Direct Probabilities
Conjugate priors

But...

Example

When $x\sim\mathcal{B}e(\alpha,\theta)$ with known α,

$$f(x|\theta)\propto\frac{\Gamma(\alpha+\theta)(1-x)^\theta}{\Gamma(\theta)},$$

conjugate distribution not so easily manageable

$$\pi(\theta|x_0,\lambda)\propto\left(\frac{\Gamma(\alpha+\theta)}{\Gamma(\theta)}\right)^\lambda(1-x_0)^\theta$$

Example

Coin spun on its edge, proportion θ of *heads*
When spinning n times a given coin, number of heads

$$x \sim \mathscr{B}(n, \theta)$$

Flat prior, or mixture prior

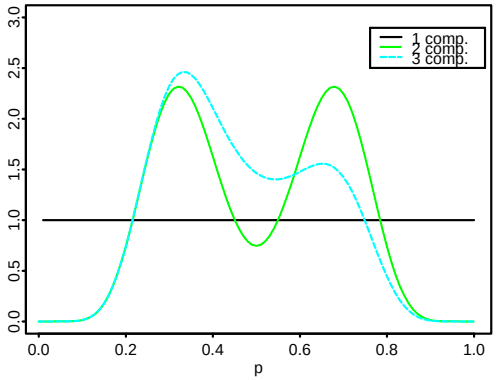
$$\frac{1}{2}\left[\mathscr{B}e(10, 20) + \mathscr{B}e(20, 10)\right]$$

or

$$0.5\,\mathscr{B}e(10, 20) + 0.2\,\mathscr{B}e(15, 15) + 0.3\,\mathscr{B}e(20, 10).$$

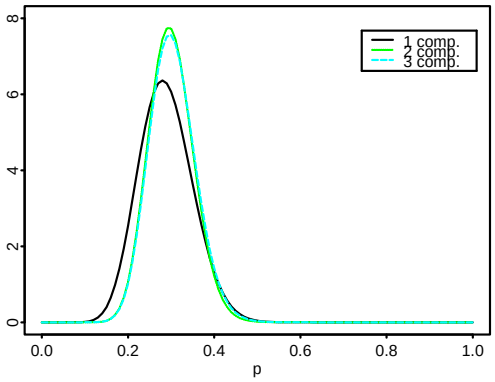
Mixtures of natural conjugate distributions also make conjugate families

Theory of Probability revisited
Direct Probabilities
Conjugate priors



Three prior distributions for a spinning-coin experiment

Theory of Probability revisited
Direct Probabilities
Conjugate priors



Posterior distributions for 50 observations

Theory of Probability revisited
Estimation problems

Chapter 3: Estimation Problems

Improper distributions

Necessary extension from a prior distribution to a prior σ-finite measure π such that

$$\int_\Theta \pi(\theta)\,d\theta = +\infty$$

...the fact that $\int_0^\infty dv/v$ diverges at both limits is a satisfactory feature (III, §3.1).

Modern justifications

Often automatic/noninformative prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior

5. Generally more acceptable to non-Bayesians, with frequentist justifications, such as:
   (i) *minimaxity*
   (ii) *admissibility*
   (iii) *invariance*
6. Improper priors prefered to vague proper priors such as a $\mathcal{N}(0, 100^2)$ distribution
7. Penalization factor in

$$\min_d \int \mathsf{L}(\theta, d)\pi(\theta) f(x|\theta)\,dx\,d\theta$$

Validation

Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior π as given by Bayes's formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

when

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty$$

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 (III, §3.1).

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

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

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

[independent of ω]

Theory of Probability revisited
Estimation problems
Improper prior distributions

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 (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)

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

...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 (III, §3.1)

[Not!]

Theory of Probability revisited
Estimation problems
Improper prior distributions

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 (III, §3.1)

Theory of Probability revisited
Estimation problems
Noninformative prior distributions

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? (III, §3.1)

Theory of Probability revisited
Estimation problems
Noninformative prior distributions

Noninformative distributions

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

Re-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]

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 (III, §3.1)

Example (Jeffreys' example, 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 properness

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

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

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 (III,§3.1)*

but so would any continuous prior!

Theory of Probability revisited
Estimation problems
Noninformative prior distributions

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 (III, §3.1)*

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

Theory of Probability revisited
Estimation problems
Noninformative prior distributions

ToP difficulties (§3.1)

End of §3.1 tries to justify the prior $\pi(v) \propto 1/v$ as "correct" prior. E.g., 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 (III, §3.1)*

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

Theory of Probability revisited
Estimation problems
Noninformative prior distributions

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 ∞ (III, §3.1)*

See also the footnote † in §3.4 !

Posterior distribution

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- extensive summary of the information available on θ
- integrate simultaneously prior information and information brought by x
- unique motor of inference

Bayesian Decision Theory

For a loss $\mathsf{L}(\theta, \delta)$ and a prior π, the *Bayes rule* is

$$\delta^\pi(x) = \arg\min_d \mathbb{E}^\pi[\mathsf{L}(\theta, d)|x].$$

Note: Practical computation not always possible analytically.

Conjugate priors

For conjugate distributions, the posterior expectations of the natural parameters can be expressed analytically, for one or several observations.

Distribution	Conjugate prior	Posterior mean
Normal	Normal	
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\dfrac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
Poisson	Gamma	
$\mathcal{P}(\theta)$	$\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$

Distribution	Conjugate prior	Posterior mean
Gamma	Gamma	
$\mathcal{G}(\nu, \theta)$	$\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + \nu}{\beta + x}$
Binomial	Beta	
$\mathcal{B}(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + n}$
Negative binomial	Beta	
$\mathcal{N}eg(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + n}{\alpha + \beta + x + n}$
Multinomial	Dirichlet	
$\mathcal{M}_k(n; \theta_1, \ldots, \theta_k)$	$\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\dfrac{\alpha_i + x_i}{\left(\sum_j \alpha_j\right) + n}$
Normal	Gamma	
$\mathcal{N}(\mu, 1/\theta)$	$\mathcal{G}(\alpha/2, \beta/2)$	$\dfrac{\alpha + 1}{\beta + (\mu - x)^2}$

Example

Consider

$$x_1, \ ..., \ x_n \sim \mathcal{U}([0, \theta])$$

and $\theta \sim \mathcal{P}a(\theta_0, \alpha)$. Then

$$\theta | x_1, ..., x_n \sim \mathcal{P}a(\max(\theta_0, x_1, ..., x_n), \alpha + n)$$

and

$$\delta^\pi(x_1, ..., x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\theta_0, x_1, ..., x_n).$$

Prediction

If $x \sim f(x|\theta)$ and $z \sim g(z|x, \theta)$, the *predictive* of z is

$$g^\pi(z|x) = \int_\Theta g(z|x, \theta)\pi(\theta|x) \, d\theta.$$

Example (AR model)

Consider the AR(1) model

$$x_t = \varrho x_{t-1} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

the predictive of x_T is then

$$x_T | x_{1:(T-1)} \sim \int \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\{-\frac{(x_T - \varrho x_{T-1})^2}{2\sigma^2}\} \pi(\varrho, \sigma | x_{1:(T-1)}) d\varrho d\sigma \, ,$$

and $\pi(\varrho, \sigma | x_{1:(T-1)})$ can be expressed in closed form

Hypergeometric and binomial inference

Case of an $\mathscr{H}(N, n, r)$ distribution under uniform prior $\pi(r) = 1/(N + 1)$
Posterior

$$P(r|N, l, H) = \binom{r}{l} \binom{N - r}{n - l} \Big/ \binom{N + 1}{n + 1}$$

Darroch model for capture–recapture

Alternative formulation:

$$n_{11} \sim \mathscr{H}(N, n_2, n_1/N)$$

Classical (MLE) estimator of N

$$\hat{N} = \frac{n_1}{(n_{11}/n_2)}$$

It cannot be used when $n_{11} = 0$

Example (Deers)

Herd of deer on an island of Newfoundland (Canada) w/o any predator, thus culling necessary for ecological equilibrium.
Annual census too time-consuming, but birth and death patterns for the deer imply that the number of deer varies between 36 and 50. Prior:

$$N \sim \mathscr{U}(\{36, \ldots, 50\})$$

Theory of Probability revisited
Estimation problems
Sampling models

Example (Deers (2))

Posterior distribution

$$\pi(N = n|n_{11}) = \frac{\binom{n_1}{n_{11}}\binom{n_2}{n_2 - n_{11}}\Big/\binom{n}{n_2}\pi(N = n)}{\sum_{k=36}^{50}\binom{n_1}{n_{11}}\binom{n_2}{n_2 - n_{11}}\Big/\binom{k}{n_2}\pi(N = k)},$$

Table: Posterior distribution of the deer population size, $\pi(N|n_{11})$.

$N \backslash n_{11}$	0	1	2	3	4	5
36	0.058	0.072	0.089	0.106	0.125	0.144
37	0.059	0.072	0.085	0.098	0.111	0.124
38	0.061	0.071	0.081	0.090	0.100	0.108
39	0.062	0.070	0.077	0.084	0.089	0.094
40	0.063	0.069	0.074	0.078	0.081	0.082
41	0.065	0.068	0.071	0.072	0.073	0.072
42	0.066	0.068	0.067	0.067	0.066	0.064
43	0.067	0.067	0.065	0.063	0.060	0.056
44	0.068	0.066	0.062	0.059	0.054	0.050
45	0.069	0.065	0.060	0.055	0.050	0.044
46	0.070	0.064	0.058	0.051	0.045	0.040
47	0.071	0.063	0.056	0.048	0.041	0.035
48	0.072	0.063	0.054	0.045	0.038	0.032
49	0.073	0.062	0.052	0.043	0.035	0.028
50	0.074	0.061	0.050	0.040	0.032	0.026

Table: Posterior mean of the deer population size, N.

n_{11}	0	1	2	3	4	5
$\mathbb{E}(N\|n_{11})$	43.32	42.77	42.23	41.71	41.23	40.78

Theory of Probability revisited
Estimation problems
Sampling models

Different loss function

$$\mathsf{L}(N,\delta) = \begin{cases} 10(\delta - N) & \text{if } \delta > N, \\ N - \delta & \text{otherwise,} \end{cases}$$

in order to avoid overestimation
Bayes estimator is $(1/11)$-quantile of $\pi(N|n_{11})$,

Table: Estimated deer population

n_{11}	0	1	2	3	4	5
$\delta^\pi(n_{11})$	37	37	37	36	36	36

Theory of Probability revisited
Estimation problems
Sampling models

Laplace succession rule

Example of a predictive distribution

> *...considering the probability that the* next *specimen will be of the first type. The population being of number N, of which n have already been removed, and the members of the first type being r in number, of which l have been removed, the probability that the next would be of the type, given r, N and the sample is (III, §3.2)*

$$P(p|l,m,N,r,H) = \frac{r-l}{N-m}\,.$$

Theory of Probability revisited
Estimation problems
Sampling models

Integrating in r

$$P(r,p|l,m,N,H) = \frac{r-l}{N-m}\binom{r}{l}\binom{N-r}{n-l}\Big/\binom{N+1}{n+1},$$

the marginal posterior of p is

$$P(p|l,m,N,H) = \frac{l+1}{N-m}\,\frac{\binom{N+1}{n+2}}{\binom{N+1}{n+1}} = \frac{l+1}{n+2}$$

> *which is independent of N. (...) Neither Bayes nor Laplace, however, seem to have considered the case of finite N (III, §3.2)*

[Why Bayes???]

New criticism of 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. (III, §3.21)*

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

Theory of Probability revisited
Estimation problems
Sampling models

Local resolution

Different weight on the boundaries

$$P(r=0|NH) = P(r=N|NH) = k$$

> *...we are therefore restricted to values of k between $\frac{1}{3}$ and $\frac{1}{2}$. A possible alternative form would be to take*
>
> $$k = \frac{1}{4} + \frac{1}{2(N+1)}$$
>
> *which puts half the prior probability into the extremes and leaves the other half distributed over all values (III, §3.21)*

Theory of Probability revisited
Estimation problems
Sampling models

Another contradiction

For the multinomial model

$$\mathscr{M}_r(n; p_1, \ldots, p_r)\,,$$

under the uniform prior

$$(p_1, \ldots, p_r) \sim \mathscr{D}(1, \ldots, 1)\,,$$

the marginal on p_1 is *not* uniform:

$$p_1 \sim \mathscr{B}(1, r-1)\,.$$

> *This expresses the fact that the average value of all the p's is not $1/r$ instead of $1/2$ (III, §3.23)*

Theory of Probability revisited
Estimation problems
Sampling models

The Poisson model

For $m \sim \mathcal{P}(r)$, Jeffreys justifies the prior $P(dr|H) \propto dr/r$ by

> *This parameter is not a chance but a chance per unit time, and therefore is dimensional (III, §3.3)*

Posterior distribution conditional on observations $m_1, \ldots, m_n$

$$P(dr|m_1, \ldots, m_n, H) \propto \frac{n^{S_m}}{(S_m-1)!}\, r^{S_m-1}\, e^{nr}\, dr$$

> *given by the incomplete Γ function. We notice that the only function of the observations that appears in the posterior probability is S_m, therefore a sufficient statistic for r (III, §3.3)*

The normal model

Importance of the normal model in many fields

$$\mathcal{N}_p(\theta, \Sigma)$$

with known Σ, normal conjugate distribution, $\mathcal{N}_p(\mu, A)$.
Under quadratic loss, the Bayes estimator is

$$\begin{aligned}\delta^\pi(x) &= x - \Sigma(\Sigma + A)^{-1}(x - \mu) \\ &= \left(\Sigma^{-1} + A^{-1}\right)^{-1}\left(\Sigma^{-1}x + A^{-1}\mu\right);\end{aligned}$$

Theory of Probability revisited
Estimation problems
Normal models and linear regression

Laplace approximation

ToP presents the normal distribution as a second order approximation of slowly varying densities,

$$P(dx|x_1, \ldots, x_n, H) \propto f(x) \exp\left\{-\frac{n}{2\sigma^2}(x - \bar{x})^2\right\}$$

(with the weird convention that $\bar{x}$ is the empirical mean of the x_i's and x is the true mean, i.e. θ...)

Theory of Probability revisited
Estimation problems
Normal models and linear regression

Estimation of variance

If

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

the likelihood is

$$\ell(\theta, \sigma \,|\, \bar{x}, s^2) \propto \sigma^{-n} \exp\left[-\frac{n}{2\sigma^2}\left\{s^2 + (\bar{x} - \theta)^2\right\}\right]$$

Jeffreys then argues in favour of

$$\pi^*(\theta, \sigma) = 1/\sigma$$

assuming *independence* between θ and σ **[Warnin!]**

Theory of Probability revisited
Estimation problems
Normal models and linear regression

In this case, the posterior distribution of (θ, σ) is such that

$$\begin{aligned}\theta|\sigma, \bar{x}, s^2 &\sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right), \\ \theta|\bar{x}, s^2 &\sim \mathcal{T}\left([n-1], \bar{x}, ns^2/[n-1]\right) \\ \sigma^2|\bar{x}, s^2 &\sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{ns^2}{2}\right).\end{aligned}$$

Reminder

- Not defined for $n = 1, 2$
- θ and σ^2 are not a posteriori independent.
- Conjugate posterior distributions have the same form
- but require a careful determination of the hyperparameters

More of the weird stuff!

Jeffreys also considers degenerate cases:

If $n = 1, \bar{x} = x_1$, and $s = 0$ **[!!!]**, *then*

$$P(dxd\sigma|x_1, H) \propto \sigma^{-2} \exp\left\{\frac{(x-\bar{x})^2}{\sigma^2}\right\} dxd\sigma$$

Integrating with respect to σ we get

$$P(dx|x_1, H) \propto \frac{dx}{|x - x_1|}$$

that is, the most probable value of x is x_1 but we have no information about the accuracy of the determination (III, §3.41).

...even though $P(dx|x_1, H)$ is not integrable...

Theory of Probability revisited
Estimation problems
Normal models and linear regression

Least squares

Usual regression model

$$y = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathsf{I}),\ \beta \in \mathbb{R}^m$$

Incredibly convoluted derivation of $\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$ in **ToP** **[see §3.5 till the normal equations in (34)]** for lack of matricial notations, replaced with tensorial conventions used by Physicists

Personally I find that to get the right value for a determinant above the third order is usually beyond my powers (III, §3.5)

...understandable in 1939 (?)...

Theory of Probability revisited
Estimation problems
Normal models and linear regression

(Modern) basics

The least-squares estimator $\hat{\beta}$ has a normal distribution

$$\hat{\beta} \sim \mathcal{N}_m(\beta, \sigma^2(X^{\mathsf{T}}X)^{-1})$$

Corresponding (Zellner's) conjugate distributions on (β, σ^2)

$$\begin{aligned} \beta|\sigma^2 &\sim \mathcal{N}_m\left(\mu, \frac{\sigma^2}{n_0}(X^{\mathsf{T}}X)^{-1}\right), \\ \sigma^2 &\sim \mathcal{IG}(\nu/2, s_0^2/2) \end{aligned}$$

Theory of Probability revisited
Estimation problems
Normal models and linear regression

since, if $s^2 = ||y - X\hat{\beta}||^2$,

$$\begin{aligned} \beta|\hat{\beta}, s^2, \sigma^2 &\sim \mathcal{N}_p\left(\frac{n_0\mu + \hat{\beta}}{n_0 + 1}, \frac{\sigma^2}{n_0 + 1}(X^{\mathsf{T}}X)^{-1}\right), \\ \sigma^2|\hat{\beta}, s^2 &\sim \mathcal{IG}\left(\frac{k - p + \nu}{2}, \frac{s^2 + s_0^2 + \frac{n_0}{n_0+1}(\mu - \hat{\beta})^{\mathsf{T}}X^{\mathsf{T}}X(\mu - \hat{\beta})}{2}\right) \end{aligned}$$

More general conjugate distributions of the type

$$\beta \sim \mathcal{N}_m(A\theta, C),$$

where $\theta \in \mathbb{R}^q$ $(q \leq m)$.

Prior modelling

In **ToP**

$$P(dx_1, \ldots, dx_m, d\sigma|H) \propto dx_1 \cdots dx_m d\sigma/\sigma$$

i.e. $\pi(\beta, \sigma) = 1/\sigma$

...the posterior probability of ζ_m is distributed as for t with $n - m$ degrees of freedom (III, §3.5)

Explanation

- the ζ_i's are the transforms of the β_i's in the eigenbasis of $(X^{\mathsf{T}}X)$
- ζ_i is also distributed as a t with $n - m$ degrees of freedom

Σ unknown

In this general case, the (apocryphal) Jeffreys prior is

$$\pi^J(\beta, \Sigma) = \frac{1}{|\Sigma|^{(k+1)/2}}.$$

with likelihood

$$\ell(\beta, \Sigma|y) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(y_i - X_i\beta)(y_i - X_i\beta)^{\mathsf{T}}\right]\right\}$$

Consequences

- suggests (inverse) Wishart distribution on Σ
- posterior marginal distribution on β only defined for sample size large enough
- no closed form expression for posterior marginal

Plusses of the Bayesian approach

Example (Truncated normal)

When $a_1, \ldots, a_n$ are $\mathcal{N}(\alpha, \sigma_r^2)$, with σ_r^2 known, and $\alpha > 0$,

the posterior probability of α is therefore a normal one about the weighted mean by the a_r, but it is truncated at $\alpha = 0$ (III,§3.55).

Separation of likelihood (observations) from prior ($\alpha > 0$)

Preliminary example

Case of a quasi-exponential setting:

$$x_1, \ldots, x_n \sim \mathcal{U}(\alpha - \sigma, \alpha + \sigma)$$

Under prior

$$P(d\alpha, d\sigma|H) \propto d\alpha\, d\sigma/\sigma$$

the two extreme observations are sufficient statistics for α and σ. Then

$$P(d\alpha|x_1, \ldots, x_n, H) \propto \begin{cases} (\alpha - x_{(1)})^{-n} d\alpha & (\alpha > (x_{(1)} + x_{(2)})/2, \\ (x_{(2)} - \alpha)^{-n} d\alpha & (\alpha < (x_{(1)} + x_{(2)})/2, \end{cases}$$

[with] a sharp peak at the mean of the extreme values (III, §3.6)

Theory of Probability revisited
Estimation problems
More sufficiency

Reassessment of sufficiency and Pitman–Koopman lemma

- in **ToP**, sufficiency is defined via a poor man's factorisation theorem, rather than through Fisher's conditional property (§3.7)
- Pitman–Koopman lemma is re-demonstrated while no mention is made of the support being independent of the parameter(s)...
- ...but Jeffreys concludes with an example where the support is $(\alpha, +\infty)$ (!)

Theory of Probability revisited
Estimation problems
More sufficiency

Partial sufficiency

Example (Normal correlation)

Case of a $\mathcal{N}_2(\theta, \Sigma)$ sample under the prior

$$\pi(\theta, \Sigma) = 1/\sigma_{11}\sigma_{22} \quad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{bmatrix}$$

Then (III, §3.9) $\pi(\sigma_{11}, \sigma_{22}, \rho|\text{Data})$ is proportional to

$$\frac{1}{(\sigma_{11}\sigma_{22})^n (1-\rho^2)^{(n-1)/2}} \exp\left\{ \frac{-n}{2(1-\rho^2)} \left(\frac{s^2}{\sigma_{11}^2} + \frac{t^2}{\sigma_{22}^2} - \frac{2\rho rst}{\sigma_{11}\sigma_{22}} \right)^2 \right\}$$

Theory of Probability revisited
Estimation problems
More sufficiency

Example (Normal correlation (2))

and

$$\pi(\rho|\text{Data}) \propto \frac{(1-\rho^2)^{(n-1)/2}}{(1-\rho r)^{n-3/2}} S_{n-1}(\rho r)$$

only depends on r, which amounts to an additional proof that r is a sufficient statistic for ρ (III, §3.9)

...Jeffreys unaware of marginalisation paradoxes...

Marginalisation paradoxes

In the case of correlation, posterior on ρ could have been derived from prior $\pi(\rho) = 1/2$ and distribution of r.
This is not always the case:

Marginalisation paradox

- $\pi(\theta_1|x_1, x_2)$ only depends on x_1
- $f(x_1|\theta_1, \theta_2)$ only depends on θ_1
- ...but $\pi(\theta_1|x_1, x_2)$ is not the same as

$$\pi(\theta_1|x_1) \propto \pi(\theta_1) f(x_1|\theta_1) \quad (!)$$

[Dawid, Stone & Zidek, 1973]

Theory of Probability revisited
Estimation problems
More sufficiency

Example (Normal MP)

Case when

$$u_1 \sim \mathcal{N}(\mu_1, \sigma^2), \quad u_2 \sim \mathcal{N}(\mu_2, \sigma^2), \quad s^2 \sim \sigma^2 \chi^2_\nu / \nu,$$

and when $\zeta = (\mu_1 - \mu_2)/(\sigma\sqrt{2})$ parameter of interest, under prior $\pi(\mu_1, \mu_2, \sigma) = 1/\sigma$
Then

- $\pi(\zeta|x)$ only depends on $z = u_1 - u_2/s\sqrt{2}$
- and z only depends on ζ
- ...but impossible to derive $\pi(\zeta|x)$ from $f(z|\zeta)$
- ...and no paradox when $\pi(\mu_1, \mu_2, \sigma) = 1/\sigma^2$ **[!!]**

(Modern, not **ToP**) invariant priors

Principle: Agree with the natural symmetries of the problem

- Identify invariance structures as group action

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

- Determine an invariant prior

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

Theory of Probability revisited
Estimation problems
More noninformative priors

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

Opening towards left- and right-Haar measures at the end of §3.10.

Invariant divergences

Interesting point made by Jeffreys that both

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

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

Theory of Probability revisited
Estimation problems
More noninformative priors

Intrinsic losses

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

[Ultimate invariance!]

Principle

Corresponding parameterisation-free loss functions:

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

Theory of Probability revisited
Estimation problems
More noninformative priors

Examples:

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

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

2. the *Hellinger distance*

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

Theory of Probability revisited
Estimation problems
More noninformative priors

Example (Normal mean)

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

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

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

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

in both cases.

Example (Normal everything)

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

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

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

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

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 hid his prior?!

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

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

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 $m \neq 2$. (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 (III, §3.10)

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

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Jeffreys' objections

Example (Normal everything)

In the case of a normal $\mathcal{N}(\lambda, \sigma^2)$, $|I(\theta)|^{1/2} = 1/\sigma^2$ instead of the prior $\pi(\theta) = 1/\sigma$ advocated earlier:

> *If the same method was applied to a joint distribution for serveral variables about independent true values, an extra factor $1/\sigma$ would appear for each. This is unacceptable: (...) λ and σ are each capable of any value over a considerable range and neither gives any appreciable information about the other (III, §3.10)*

Example (Variable support)

If the support of $f(\cdot|\theta)$ depends on θ, e.g. $\mathcal{U}([\theta_1, \theta_2])$ usually no Fisher information [e.g. exception $\propto \{(x-\theta_1)^+\}^a\{(\theta_2 - x)^+\}^b$ with $a, b > 1$]. Jeffreys suggests to condition on non-differentiable parameters to derive a prior on the other parameters, and to use a flat prior on the bounds of the support.

Theory of Probability revisited
Estimation problems
The Jeffreys prior
Theory of Probability revisited
Estimation problems
The Jeffreys prior

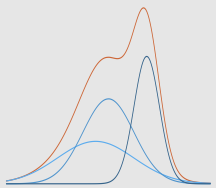
Poisson distribution

> *The Poisson parameter, however, is in rather a special position. It is usually the product of a scale factor with an arbitrary sample size, which is not chosen until we have already have some information about the probable range of values for the scale parameter. It does however point a warning for all designed experiments. The whole point of general rules for the prior probability is to give a starting-point, which we take to represent ignorance. They will not be correct if previous knowledge is being used (...) In the case of the Poisson law the sample size is chosen so that λ will be a moderate number, usually 1 to 10. The $d\lambda/\lambda$ rule, in fact, may express complete ignorance of the scale parameter; but $d\lambda/\sqrt{\lambda}$ may express just enough information to suggest that the experiment is worth making.*

Example (Mixture model)

For

$$f(x|\theta) = \sum_{i=1}^{k} \omega_i f_i(x|\alpha_i)\,,$$



Jeffreys suggests to separate the ω_i's from the α_i's:

$$\pi^J(\omega, \alpha) \propto \prod_{i=1}^{k} |I(\alpha_i)|^{1/2} / \sqrt{\omega_i} \tag{36}$$

Exponential families

Jeffreys makes yet another exception for *Huzurbazar* distributions

$$f(x) = \phi(\alpha)\psi(x)\exp\{u(\alpha)v(x)\}$$

namely exponential families.
Using the reparameterisation $\beta = u(\alpha)$, he considers three cases

1. $\beta \in (-\infty, +\infty)$, then $\pi^\star(\beta) \propto 1$
2. $\beta \in (0, +\infty)$, then $\pi^\star(\beta) \propto 1/\beta$
3. $\beta \in (0, 1)$, then $\pi^\star(\beta) = 1/\beta(1-\beta)$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Pros & Cons

- Parameterization invariant
- Relates to information theory
- Agrees with most invariant priors (e.g., location/scale)
- Suffers from dimensionality curse (e.g., Jeffreys' correction)
- Not coherent for Likelihood Principle (e.g., Binomial versus Negative binomial)

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Example (Normal norm)

$$x \sim \mathcal{N}_p(\theta, I_p), \qquad \eta = \|\theta\|^2, \qquad \pi(\eta) = \eta^{p/2-1}$$
$$\mathbb{E}^\pi[\eta|x] = \|x\|^2 + p \qquad \text{Bias } 2p$$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Example (Likelihood paradox)

If $x \sim \mathscr{B}(n, \theta)$, Jeffreys' prior is

$$\mathscr{B}e(1/2, 1/2)$$

and, if $n \sim \mathscr{N}eg(x, \theta)$, Jeffreys' prior is

$$\begin{aligned}\pi_2(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] \\ &= \mathbb{E}_\theta\left[\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right] = \frac{x}{\theta^2(1-\theta)}, \\ &\propto \theta^{-1}(1-\theta)^{-1/2}\end{aligned}$$

Bernardo's reference priors

Generalizes Jeffreys priors by distinguishing between nuisance and interest parameters
Principle: maximize the information brought by the data

$$\mathbb{E}^n \left[\int \pi(\theta|x_n) \log(\pi(\theta|x_n)/\pi(\theta)) d\theta \right]$$

and consider the limit of the π_n
Outcome: most usually, Jeffreys prior

Nuisance parameters

For $\theta = (\lambda, \omega)$,

$$\pi(\lambda|\omega) = \pi_J(\lambda|\omega) \qquad \text{with fixed } \omega$$

Jeffreys' prior conditional on ω, and

$$\pi(\omega) = \pi_J(\omega)$$

for the marginal model

$$f(x|\omega) \propto \int f(x|\theta)\pi_J(\lambda|\omega)d\lambda$$

- Depends on ordering
- Problems of definition

Example (Neyman–Scott problem)

Observation of x_{ij} iid $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$.
The usual Jeffreys prior for this model is

$$\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-n-1}$$

which is inconsistent because

$$\mathbb{E}[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(2n-2),$$

where

$$s^2 = \sum_{i=1}^{n} \frac{(x_{i1} - x_{i2})^2}{2},$$

Example (Neyman–Scott problem (2))

Associated reference prior with $\theta_1 = \sigma$ and $\theta_2 = (\mu_1, \ldots, \mu_n)$ gives

$$\begin{aligned} \pi(\theta_2|\theta_1) &\propto 1\,, \\ \pi(\sigma) &\propto 1/\sigma \end{aligned}$$

Therefore,

$$\mathbb{E}[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(n-2)$$

Matching priors

Frequency-validated priors:
Some posterior probabilities

$$\pi(g(\theta) \in C_x|x) = 1-\alpha$$

must coincide with the corresponding frequentist coverage

$$P_\theta(C_x \ni g(\theta)) = \int \mathbb{I}_{C_x}(g(\theta))\, f(x|\theta)\, dx\,,$$

...asymptotically

Theory of Probability revisited
Estimation problems
The Jeffreys prior

For instance, Welch and Peers' identity

$$P_\theta(\theta \le k_\alpha(x)) = 1-\alpha + O(n^{-1/2})$$

and for Jeffreys' prior,

$$P_\theta(\theta \le k_\alpha(x)) = 1-\alpha + O(n^{-1})$$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

In general, choice of a matching prior dictated by the cancelation of a first order term in an **Edgeworth expansion**, like

$$[I''(\theta)]^{-1/2} I'(\theta) \nabla \log \pi(\theta) + \nabla^{\mathsf{T}} \{I'(\theta)[I''(\theta)]^{-1/2}\} = 0\,.$$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Example (Linear calibration)

$(i = 1, \ldots, n,\ j = 1, \ldots, k)$

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad y_{0j} = \alpha + \beta x_0 + \varepsilon_{0j}\,,$$

with $\theta = (x_0, \alpha, \beta, \sigma^2)$ and x_0 quantity of interest

Example (Linear calibration (2))

One-sided differential equation:

$$|\beta|^{-1}s^{-1/2}\frac{\partial}{\partial x_0}\{e(x_0)\pi(\theta)\} - e^{-1/2}(x_0)\text{sgn}(\beta)n^{-1}s^{1/2}\frac{\partial\pi(\theta)}{\partial x_0}$$
$$-e^{-1/2}(x_0)(x_0-\bar{x})s^{-1/2}\frac{\partial}{\partial\beta}\{\text{sgn}(\beta)\pi(\theta)\} = 0$$

with

$$s = \Sigma(x_i-\bar{x})^2,\ \ e(x_0) = [(n+k)s + nk(x_0-\bar{x})^2]/nk\,.$$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Example (Linear calibration (3))

Solutions

$$\pi(x_0,\alpha,\beta,\sigma^2) \propto e(x_0)^{(d-1)/2}|\beta|^d g(\sigma^2)\,,$$

where g arbitrary.

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Reference priors

Partition	Prior
$(x_0,\alpha,\beta,\sigma^2)$	$\|\beta\|(\sigma^2)^{-5/2}$
$x_0,\alpha,\beta,\sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0,\alpha,(\sigma^2,\beta)$	$e(x_0)^{-1/2}(\sigma^2)^{-3/2}$
$x_0,(\alpha,\beta),\sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0,(\alpha,\beta,\sigma^2)$	$e(x_0)^{-1/2}(\sigma^2)^{-2}$

Theory of Probability revisited
Estimation problems
The Jeffreys prior

Other approaches

- Rissanen's transmission information theory and minimum length priors
- Testing priors
- stochastic complexity

Fourth chapter: Approximate methods and simplifications

MAP

Equivalence of MAP and ML estimators:

> *...the differences between the values that make the likelihood and the posterior density maxima are only of order $1/n$ (IV, §4.0)*

extrapolated into

> *...in the great bulk of cases the results of [the method of maximum likelihood] are undistinguishable from those given by the principle of inverse probability (IV, §4.0)*

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? (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

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

Therefore posterior median is $2m$

No mention made of either MLE or unbiasedness

Laplace analytic approximation

When integrating a regular function

$$\mathbb{E}^\pi[g(\theta)|x] = \frac{\int_\Theta g(\theta)f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta} = \frac{\int_\Theta b_N(\theta)\exp\{-nh_N(\theta)\}\,d\theta}{\int_\Theta b_D(\theta)\exp\{-nh_D(\theta)\}\,d\theta},$$

Laplace's approximation given by

$$\frac{\int_\Theta b_N(\theta)\exp\{-nh_N(\theta)\}\,d\theta}{\int_\Theta b_D(\theta)\exp\{-nh_D(\theta)\}\,d\theta} = \frac{\sigma_N}{\sigma_D}e^{-n(\hat{h}_N-\hat{h}_D)}\left[\frac{\hat{b}_N}{\hat{b}_D}+\frac{\sigma_D^2}{2n\hat{b}_D^2}\left\{\hat{b}_D\hat{b}_N''\right.\right.$$
$$\left.\left.-\hat{b}_N\hat{b}_D''-\sigma_D^2\hat{h}_D'''(\hat{b}_D\hat{b}_N'-\hat{b}_N\hat{b}_D')\right\}\right]+\mathsf{O}(n^{-2}).$$

Consequence

$$\mathbb{E}^\pi[g(\theta)|x] = \hat{g}+\frac{\sigma_D^2\hat{b}_D'\hat{g}'}{n\hat{b}_D}+\frac{\sigma_D^2\hat{g}''}{2n}-\frac{\sigma_D^4\hat{h}'''\hat{g}'}{2n}+\mathsf{O}(n^{-2}).$$

Example (Binomial model)

$\pi(\theta|x)$ density of $\mathscr{B}e(\alpha,\beta)$ distribution and posterior expectation of θ

$$\delta^\pi(x) = \frac{\alpha}{\alpha+\beta},$$

compared with

$$\delta^\pi(x) = \frac{\alpha^2+\alpha\beta+2-4\alpha}{(\alpha+\beta-2)^2}+\mathsf{O}((\alpha+\beta)^{-2}),$$

Fighting un-sufficiency

When maximum likelihood estimators not easily computed (e.g., outside exponential families), Jeffreys suggests use of Pearson's *minimum* χ^2 estimation, which is a form of MLE for multinomial settings.

Asymptotic difficulties of

> *In practice, (...) it is enough to group [observations] so that there are no empty groups, m_r for a terminal group being calculated for a range extending to infinity (IV, §4.1)*

bypassed in **ToP**

Unbiasedness

Searching for unbiased estimators presented in §4.3 as a way of fighting un-sufficiency and attributed to Neyman and Pearson.

Introduction of Decision Theory via a multidimensional loss function:

> *There are apparently an infinite number of unbiased statistics associated with any law (...) The estimates of $\alpha,\beta,\ldots$ obtained will therefore be $a,b,\ldots$ which differ little from $\alpha,\beta,\ldots$ The choice is then made so that all of $E(\alpha-a)^2, E(\beta-b)^2,\ldots$ will be as small as possible*

Note the first sentence above: **meaningless!**
Besides, unbiasedness is a property **almost never** shared by Bayesian estimators!

Evaluating estimators

Purpose of most inferential studies

To provide the statistician/client with a *decision* $d \in \mathscr{D}$
Requires an evaluation criterion for decisions and estimators

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

[a.k.a. loss function]

Bayesian Decision Theory

Three spaces/factors:

(1) On $\mathscr{X}$, distribution for the observation, $f(x|\theta)$;
(2) On Θ, prior distribution for the parameter, $\pi(\theta)$;
(3) On $\Theta \times \mathscr{D}$, loss function associated with the decisions, $\mathsf{L}(\theta, \delta)$;

Foundations

Theorem (**Existence**)

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

[DeGroot, 1970]

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

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

when θ is unknown

Frequentist Principle

Average loss (or frequentist risk)

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

Principle

Select the best estimator based on the risk function

Difficulties with frequentist paradigm

(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!

(2) Assumption of repeatability of experiments not always grounded.

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

Bayesian principle

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

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

Bayesian principle (2)

Alternative

Integrate over the space Θ and compute *integrated risk*

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

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

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

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*

Infinite Bayes risk

Above result valid for both proper and improper priors when

$$r(\pi) < \infty$$

Otherwise, **generalized Bayes estimator** that must be defined pointwise:

$$\delta^\pi(x) = \arg\min_{d} \rho(\pi, d|x)$$

if $\rho(\pi, d|x)$ is well-defined for every x.

Warning: Generalized Bayes $\neq$ Improper Bayes

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

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

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

  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

Example (Normal mean)

Consider $x \sim \mathcal{N}(\theta, 1)$ and the test of $H_0 : \theta \leq 0$, i.e. the estimation of

$$\mathbb{I}_{H_0}(\theta)$$

Under the loss

$$(\mathbb{I}_{H_0}(\theta) - \delta(x))^2\,,$$

the estimator (p-value)

$$\begin{aligned} p(x) &= P_0(X > x) \qquad (X \sim \mathcal{N}(0,1)) \\ &= 1 - \Phi(x), \end{aligned}$$

is Bayes under Lebesgue measure.

Theory of Probability revisited
Asymptotics & DT& ...
Admissibility

Example (Normal mean (2))

Indeed

$$\begin{aligned} p(x) &= \mathbb{E}^\pi[\mathbb{I}_{H_0}(\theta)|x] = P^\pi(\theta < 0|x) \\ &= P^\pi(\theta - x < -x|x) = 1 - \Phi(x). \end{aligned}$$

The Bayes risk of p is finite and $p(s)$ is **admissible**.

Example (Normal mean (3))

Consider $x \sim \mathcal{N}(\theta, 1)$. Then $\delta_0(x) = x$ is a generalised Bayes estimator, is admissible, but

$$\begin{aligned} r(\pi, \delta_0) &= \int_{-\infty}^{+\infty} R(\theta, \delta_0)\, d\theta \\ &= \int_{-\infty}^{+\infty} 1\, d\theta = +\infty. \end{aligned}$$

Example (Normal mean (4))

Consider $x \sim \mathcal{N}_p(\theta, I_p)$. If

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

the Bayes estimator for the Lebesgue measure is

$$\delta^\pi(x) = ||x||^2 + p.$$

This estimator is not admissible because it is dominated by

$$\delta_0(x) = ||x||^2 - p$$

The quadratic loss

Historically, first loss function (Legendre, Gauss)

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

or

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

> *The reason for using the expectation of the square of the error as the criterion is that, given a large number of observations, the probability of a set of statistics given the parameters, and that of the parameters given the statistics, are usually distributed approximately on a normal correlation surface (IV, §4.3)*

Explntn: Asymptotic normal distribution of MLE

Proper loss

Posterior mean

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

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

Orthogonal parameters

Interesting disgression: reparameterise the parameter set so that Fisher information is (nearly) diagonal.

> *...the quadratic term in $E(\log L)$ will reduce to a sum of squares (IV, §4.3)*

But this is *local orthogonality*: the diagonal terms in $I(\theta)$ may still depend on all parameters and Jeffreys distinguishes *global orthogonality* where each diagonal term only depends on one β_i and thus induces an independent product for the Jeffreys prior.

Generaly impossible, even though interesting for dealing with nuisance parameters...

The absolute error loss

Alternatives to the quadratic loss:

$$\mathsf{L}(\theta, d) = |\,\theta - d\,|\,,$$

or

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

L_1 estimator

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

Posterior median

Relates to Jeffreys'

> *...we can use the median observation as a statistic for the median of the law (IV, §4.4)*

even though it lacks DT justification

Example (Median law)

If

$$P(dx|m, a, H) \propto \frac{1}{2} \exp\left(-\frac{|x - m|}{a}\right) \frac{dx}{a}$$

the likelihood is maximum if m is taken equal to the median observation and if a is the average residual without regard to sign.

['tis Laplace's law]

> *It is only subject to that law that the average residual leads to the best estimate of uncertainty, and then the best estimate of location is provided by the median observation and not by the mean (IV, §4.4)*

No trace whatsoever of Bayesian estimation???

ToP 4 bric-à-brac!

Sequence of remarks and cases

- Difficulty when no sufficient statistics
- Model/law misspecification (departure from normality)
- Random effect
- Randomization (contradiction to later Bayesian persp's)
- Rank tests (Spearman: *It is an estimate but what is it an estimate of?*)

with very little relevance to either Bayesian methodology or DT...

Maybe a reflection on computational difficulties?

Chapter 5: Significance tests: one new parameter

Theory of Probability revisited
Significance tests: one new parameter

Fundamental setting

Is the new parameter supported by the observations or is any variation expressible by it better interpreted as random? Thus we must set two hypotheses for comparison, the more complicated having the smaller initial probability (V, §5.0)

[Occam's rule again!]

...compare a specially suggested value of a new parameter, often 0 [q], with the aggregate of other possible values [q']. We shall call q the null hypothesis and q' the alternative hypothesis [and] we must take

$$P(q|H) = P(q'|H) = 1/2\,.$$

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Example (Normal mean)

For $x \sim \mathcal{N}(\theta, 1)$, decide whether or not $\theta \leq 0$.

The $0-1$ loss

Neyman–Pearson loss for testing hypotheses

Test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$.
Then

$$\mathscr{D} = \{0, 1\}$$

The $0-1$ loss

$$\mathsf{L}(\theta, d) = \begin{cases} 1-d & \text{if } \theta \in \Theta_0 \\ d & \text{otherwise,} \end{cases}$$

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Type–one and type–two errors

Associated with the risk

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathsf{L}(\theta, \delta(x))] \\ &= \begin{cases} P_\theta(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ P_\theta(\delta(x) = 1) & \text{otherwise,} \end{cases} \end{aligned}$$

Theorem (Bayes test)

The Bayes estimator associated with π and with the $0-1$ loss is

$$\delta^\pi(x) = \begin{cases} 1 & \textit{if } P(\theta \in \Theta_0 | x) > P(\theta \notin \Theta_0 | x), \\ 0 & \textit{otherwise,} \end{cases}$$

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Jeffreys' example (§5.0)

Testing whether the mean α of a normal observation is zero:

$$P(q|aH) \propto \exp\left(-\frac{a^2}{2s^2}\right)$$

$$P(q'\mathrm{d}\alpha|aH) \propto \exp\left(-\frac{(a-\alpha)^2}{2s^2}\right) f(\alpha)\mathrm{d}\alpha$$

$$P(q'|aH) \propto \int \exp\left(-\frac{(a-\alpha)^2}{2s^2}\right) f(\alpha)\mathrm{d}\alpha$$

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

A point of contention

Jeffreys asserts

> *Suppose that there is one old parameter α; the new parameter is β and is 0 on q. In q' we could replace α by α', any function of α and β: but to make it explicit that q' reduces to q when $\beta = 0$ we shall require that $\alpha' = \alpha$ when $\beta = 0$ (V, §5.0).*

This amounts to assume identical parameters in both models, a controversial principle for model choice (see Chapter 6) or at the very best to make α and β dependent a priori, a choice contradicted by the following paragraphs!

Orthogonal parameters

If

$$I(\alpha,\beta) = \begin{bmatrix} g_{\alpha\alpha} & 0 \\ 0 & g_{\beta\beta} \end{bmatrix},$$

α and β *orthogonal*, but not [a posteriori] *independent*, contrary to **ToP** assertions

> *...the result will be nearly independent on previous information on old parameters (V, §5.01).*

and

$$K = \frac{1}{f(b,a)}\sqrt{\frac{ng_{\beta\beta}}{2\pi}}\exp\left(-\frac{1}{2}ng_{\beta\beta}b^2\right)$$

> *[where] $h(\alpha)$ is irrelevant (V, §5.01)*

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Acknowledgement in **ToP**

> *In practice it is rather unusual for a set of parameters to arise in such a way that each can be treated as irrelevant to the presence of any other. More usual cases are (...) where some parameters are so closely associated that one could hardly occur without the others (V, §5.04).*

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Generalisation

Theorem (Optimal Bayes decision)

Under the $0-1$ loss function

$$L(\theta,d) = \begin{cases} 0 & \text{if } d = \mathbb{I}_{\Theta_0}(\theta) \\ a_0 & \text{if } d = 1 \text{ and } \theta \notin \Theta_0 \\ a_1 & \text{if } d = 0 \text{ and } \theta \in \Theta_0 \end{cases}$$

the Bayes procedure is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \Pr^\pi(\theta \in \Theta_0|x) \geq a_0/(a_0+a_1) \\ 0 & \text{otherwise} \end{cases}$$

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances
Often difficult to assess in practice and replacement with "golden" bounds like .05, biased towards H_0

Example (Binomial probability)

Consider $x \sim \mathscr{B}(n,p)$ and $\Theta_0 = [0,1/2]$. Under the uniform prior $\pi(p) = 1$, the posterior probability of H_0 is

$$P^\pi(p \leq 1/2|x) = \frac{\int_0^{1/2} p^x(1-p)^{n-x}\mathrm{d}p}{B(x+1,n-x+1)}$$
$$= \frac{(1/2)^{n+1}}{B(x+1,n-x+1)}\left\{\frac{1}{x+1}+\ldots+\frac{(n-x)!x!}{(n+1)!}\right\}$$

Loss/prior duality

Decomposition

$$\begin{aligned}\Pr^\pi(\theta \in \Theta_0|x) &= \int_{\Theta_0} \pi(\theta|x)\,\mathrm{d}\theta \\ &= \frac{\int_{\Theta_0} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}{\int_{\Theta} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}\end{aligned}$$

suggests representation

$$\pi(\theta) = \pi(\Theta_0)\pi_0(\theta) + (1-\pi(\Theta_0))\pi_1(\theta)$$

and decision

$$\delta^\pi(x) = 1 \text{ iff } \frac{\pi(\Theta_0)}{(1-\pi(\Theta_0))}\frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)\,\mathrm{d}\theta}{\int_{\Theta_0^c} f(x|\theta_0)\pi_1(\theta)\,\mathrm{d}\theta} \geq \frac{a_0}{a_1}$$

©What matters is $(\pi(\Theta_0)/a_0, (1-\pi(\Theta_0))/a_1)$

A function of posterior probabilities

Definition (Bayes factors)

For hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_\mathsf{a} : \theta \notin \Theta_0$

$$B_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0^c|x)} \Big/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \frac{\displaystyle\int_{\Theta_0} f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\displaystyle\int_{\Theta_0^c} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta}$$

[Good, 1958 & **ToP**, V, §5.01]

▸ Goto Poisson example

Equivalent to Bayes rule: acceptance if $B_{01} > \{(1-\pi(\Theta_0))/a_1\}/\{\pi(\Theta_0)/a_0\}$

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio

  Jeffreys' scale of evidence (Appendix B):
  - if $\log_{10}(B_{10}^\pi) < 0$ null H_0 *supported*
  - if $\log_{10}(B_{10}^\pi)$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(B_{10}^\pi)$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B_{10}^\pi)$ 1 and 1.5, evidence *strong*,
  - if $\log_{10}(B_{10}^\pi)$ 1.5 and 2, evidence *very strong* and
  - if $\log_{10}(B_{10}^\pi)$ above 2, evidence *decisive*

Hot hand

Example (Binomial homogeneity)

Consider $H_0 : y_i \sim \mathscr{B}(n_i, p)$ $(i = 1, \ldots, G)$ vs. $H_1 : y_i \sim \mathscr{B}(n_i, p_i)$. Conjugate priors $p_i \sim \mathscr{B}e(\alpha = \xi/\omega, \beta = (1-\xi)/\omega)$, with a uniform prior on $\mathbb{E}[p_i|\xi,\omega] = \xi$ and on p (ω is fixed)

$$\begin{aligned}B_{10} = &\int_0^1 \prod_{i=1}^G \int_0^1 p_i^{y_i}(1-p_i)^{n_i-y_i} p_i^{\alpha-1}(1-p_i)^{\beta-1}\mathrm{d}\,p_i \\ &\times\Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)]\mathrm{d}\xi \\ &\Big/ \int_0^1 p^{\sum_i y_i}(1-p)^{\sum_i (n_i-y_i)}\mathrm{d}\,p\end{aligned}$$

For instance, $\log_{10}(B_{10}) = -0.79$ for $\omega = 0.005$ and $G = 138$ slightly favours H_0.

[Kass & Raftery, 1995]

Multiple alternatives (§5.03)

If $q' = q_1 \cup \cdots \cup q_m$, and if $P(q'|H) = 1/2$, then, if $P(q_i|H) = \kappa$,

$$(1-\kappa)^m = \frac{1}{2}$$

and

$$\frac{P(q|H)}{P(q_i|H)} \approx \frac{m}{2\log 2} = 0.7m$$

© **If testing for a separate hypothesis q_i, Bayes factor B_{0i} multiplied by $0.7m$**

A major modification

When the null hypothesis is supported by a set of measure 0, $\pi(\Theta_0) = 0$

[End of the story?!]

Suppose we are considering whether a location parameter α is 0. The estimation prior probability for it is uniform and we should have to take $f(\alpha) = 0$ and $K[= B_{10}]$ would always be infinite (V, §5.02)

Requirement

Defined prior distributions under both assumptions,

$$\pi_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad \pi_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(under the standard dominating measures on Θ_0 and Θ_1)

Using the prior probabilities $\pi(\Theta_0) = \varrho_0$ and $\pi(\Theta_1) = \varrho_1$,

$$\pi(\theta) = \varrho_0\pi_0(\theta) + \varrho_1\pi_1(\theta).$$

Note If $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass in θ_0

Contingency table (§5.11)

Then the alternatives, the sampling numbers, and the chances may be shown as follows:

$$\begin{pmatrix} \phi.\psi & \phi.\tilde{\psi} \\ \tilde{\phi}.\psi & \tilde{\phi}.\tilde{\psi} \end{pmatrix}, \quad \begin{pmatrix} x & y \\ x' & y' \end{pmatrix}, \quad \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}.$$

If ϕ and ψ are in proportion we have hypothesis q, that

$$p_{11}p_{22} = p_{12}p_{21}\,.$$

Contingency table (cont'd)

Under q,

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \alpha\beta & \alpha(1-\beta) \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) \end{pmatrix}$$

and under q',

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \alpha\beta+\gamma & \alpha(1-\beta)-\gamma \\ (1-\alpha)\beta-\gamma & (1-\alpha)(1-\beta)+\gamma \end{pmatrix}$$

If $\alpha \leq \beta \leq \frac{1}{2}$, then $-\alpha\beta \leq \gamma \leq \alpha(1-\beta)$.

Theory of Probability revisited
Significance tests: one new parameter
Bayes factors

Contingency table (cont'd)

In general, it should be

$$-\{\alpha\beta \wedge (1-\alpha)(1-\beta)\} \leq \gamma \leq \{\alpha(1-\beta) \wedge (1-\alpha)\beta\}$$

Then

$$\pi_1(\alpha,\beta,\gamma) = \frac{1}{\min(\alpha, 1-\alpha, \beta, 1-\beta)} \times \mathbb{I}_{(-(\alpha\beta\wedge(1-\alpha)(1-\beta)),(\alpha(1-\beta)\wedge(1-\alpha)\beta))}(\gamma)$$

and

$$P(q|\theta H) \propto \frac{(x+y)!(x'+y')!(x+x')!(y+y')!}{\{(x+y+x'+y')\}^2}$$

Theory of Probability revisited
Significance tests: one new parameter
Bayes factors

Contingency table (cont'd)

...but ToP gets it wrong when integrating in $P(q'|\theta H)$ since it keeps dividing by α rather than by $\min(\alpha, 1-\alpha, \beta, 1-\beta)$...

Obvious **ToP** difficulty in computing

$$\int (\alpha\beta+\gamma)^x(\alpha(1-\beta)-\gamma)^y((1-\alpha)\beta-\gamma)^{x'} ((1-\alpha)(1-\beta)+\gamma)^{y'}\pi_1(\mathrm{d}\alpha,\mathrm{d}\beta,\mathrm{d}\gamma)$$

MC resolution

1. Simulate $(\alpha,\beta,\gamma) \sim \pi_1(\alpha,\beta,\gamma)$
2. Average $(\alpha\beta+\gamma)^x(\alpha(1-\beta)-\gamma)^y((1-\alpha)\beta-\gamma)^{x'}((1-\alpha)(1-\beta)+\gamma)^{y'}$

Theory of Probability revisited
Significance tests: one new parameter
Bayes factors

A touch of Eugenics...

...data on the conviction of twin brothers or sisters (of like sex) of convicted criminals, according as the twins were monozygotic (identical) or dizygotic (no more alike physically than ordinary brothers or sisters)

	Monogozytic	Dizygotic
Convicted	*10*	*2*
Not convicted	*3*	*15*

Then

$$K = \frac{1}{171}$$

(...) we can assert on the data that the odds on the existence of a difference are about 171 *to* 1 *(V, §5.14)*

Point null hypotheses

Particular case $H_0 : \theta = \theta_0$
Take $\rho_0 = \Pr^\pi(\theta = \theta_0)$ and g_1 prior density under H_a.
Posterior probability of H_0

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta)\,d\theta} = \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + (1-\rho_0)m_1(x)}$$

and marginal under H_a

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta.$$

Point null hypotheses (cont'd)

Dual representation

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

and

$$B^\pi_{01}(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Big/ \frac{\rho_0}{1-\rho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

Connection

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{B^\pi_{01}(x)}\right]^{-1}.$$

Point null hypotheses (cont'd)

Example (Normal mean)

Test of $H_0 : \theta = 0$ when $x \sim \mathcal{N}(\theta, 1)$: we take π_1 as $\mathcal{N}(0, \tau^2)$

$$\begin{aligned}\frac{m_1(x)}{f(x|0)} &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\}\end{aligned}$$

and

$$\pi(\theta = 0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}$$

Point null hypotheses (cont'd)

Example (Normal mean)

Influence of τ:

τ/x	0	0.68	1.28	1.96
1	0.586	0.557	0.484	0.351
10	0.768	0.729	0.612	0.366

A fundamental difficulty

Improper priors are not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then either π_1 or π_2 cannot be coherently normalised **but** the normalisation matters in the Bayes factor Recall Bayes factor

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Constants matter

Example (Poisson versus Negative binomial)

If $\mathfrak{M}_1$ is a $\mathscr{P}(\lambda)$ distribution and $\mathfrak{M}_2$ is a $\mathscr{NB}(m,p)$ distribution, we can take

$$\begin{aligned} \pi_1(\lambda) &= 1/\lambda \\ \pi_2(m,p) &= \tfrac{1}{M}\,\mathbb{I}_{\{1,\cdots,M\}}(m)\,\mathbb{I}_{[0,1]}(p) \end{aligned}$$

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Constants matter (cont'd)

Example (Poisson versus Negative binomial (2))

then

$$\begin{aligned} B^{\pi}_{12} &= \frac{\displaystyle\int_0^{\infty} \frac{\lambda^{x-1}}{x!} e^{-\lambda}\mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_{m=1}^{M}\int_0^{\infty}\binom{m}{x-1}p^{x}(1-p)^{m-x}dp} \\ &= 1\Big/\frac{1}{M}\sum_{m=x}^{M}\binom{m}{x-1}\frac{x!(m-x)!}{m!} \\ &= 1\Big/\frac{1}{M}\sum_{m=x}^{M} x/(m-x+1) \end{aligned}$$

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Improper priors on common (nuisance) parameters do not matter (so much)

Normal illustration

Take $x \sim \mathcal{N}(\theta, 1)$ and $H_0 : \theta = 0$

Influence of the constant

$\pi(\theta)/x$	0.0	1.0	1.65	1.96	2.58
1	0.285	0.195	0.089	0.055	0.014
10	0.0384	0.0236	0.0101	0.00581	0.00143

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

ToP unaware of the problem?

Example of testing for a zero normal mean:

> *If σ is the standard error and λ the true value, λ is 0 on q. We want a suitable form for its prior on q'. (...) Then we should take*
>
> $$P(q\,d\sigma|H) \propto d\sigma/\sigma$$
> $$P(q'\,d\sigma\,d\lambda|H) \propto f\left(\frac{\lambda}{\sigma}\right) d\sigma/\sigma\, d\lambda/\lambda$$
>
> *where f [is a true density] (V, §5.2).*

Fallacy of the "same" σ!

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Not enought information

If $s' = 0$ [!!!], then [for $\sigma = |\bar{x}|/\tau$, $\lambda = \sigma v$]

$$P(q|\theta H) \propto \int_0^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n \exp\left(-\frac{1}{2}n\tau^2\right) \frac{d\tau}{\tau},$$
$$P(q'|\theta H) \propto \int_0^\infty \frac{d\tau}{\tau} \int_{-\infty}^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n f(v) \exp\left(-\frac{1}{2}n(v-\tau)^2\right).$$

If $n = 1$ and $f(v)$ is any even [density],

$$P(q'|\theta H) \propto \frac{1}{2}\frac{\sqrt{2\pi}}{|\bar{x}|} \quad \text{and} \quad P(q|\theta H) \propto \frac{1}{2}\frac{\sqrt{2\pi}}{|\bar{x}|}$$

and therefore $K = 1$ (V, §5.2).

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Strange constraints

> *If $n \geq 2$, the condition that $K = 0$ for $s' = 0$, $\bar{x} \neq 0$ is equivalent to*
>
> $$\int_0^\infty f(v) v^{n-1} dv = \infty\,.$$
>
> *The function satisfying this condition for [all] n is*
>
> $$f(v) = \frac{1}{\pi(1+v^2)}$$

This is the prior recommended by Jeffreys hereafter.
But, first, many other families of densities satisfy this constraint and a scale of 1 cannot be **universal**!
Second, $s' = 0$ is a zero probability event...

Further puzzlements!

When taking two normal sample $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$ with means λ_1 and λ_2 and same variance σ, testing for $H_0 : \lambda_1 = \lambda_2$ suddenly gets outwordly:

> *...we are really considering four hypotheses, not two as in the test for agreement of a location parameter with zero; for neither may be disturbed, or either, or both may.*

ToP then uses parameters (λ, σ) in all versions of the alternative hypotheses, with

$$\begin{aligned}
\pi_0(\lambda, \sigma) &\propto 1/\sigma \\
\pi_1(\lambda, \sigma, \lambda_1) &\propto 1/\pi\{\sigma^2 + (\lambda_1 - \lambda)^2\} \\
\pi_2(\lambda, \sigma, \lambda_2) &\propto 1/\pi\{\sigma^2 + (\lambda_2 - \lambda)^2\} \\
\pi_{12}(\lambda, \sigma, \lambda_1, \lambda_2) &\propto \sigma/\pi^2\{\sigma^2 + (\lambda_1 - \lambda)^2\}\{\sigma^2 + (\lambda_2 - \lambda)^2\}
\end{aligned}$$

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

ToP misses the points that

1. λ not have the same meaning under q, under q_1 $(= \lambda_2)$ and under q_2 $(= \lambda_1)$
2. λ has no precise meaning under q_{12} [hyperparameter?]

   > *On q_{12}, since λ does not appear explicitely in the likelihood we can integrate it (V, §5.41).*

3. even σ has a varying meaning over hypotheses
4. integrating over measures is meaningless!

$$P(q_{12}\mathrm{d}\sigma\mathrm{d}\lambda_1\mathrm{d}\lambda_2|H) \propto \frac{2}{\pi}\frac{\mathrm{d}\sigma\mathrm{d}\lambda_1\mathrm{d}\lambda_2}{4\sigma^2 + (\lambda_1 - \lambda_2)^2}$$

simply defines a new prior...

Further, erases the fake complexity in the end:

> *But there is so little to choose between the alternatives that we may as well combine them (V, §5.41).*

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Similar confusion in following sections (§5.42 — §5.45): the use of improper priors in testing settings simply does not make sense because ... constants matter!

Note also the aggravating effect of the multiple alternatives (e.g., §5.46):

$$P(q'|\theta H) = P(q_1|\theta H) + P(q_2|\theta H) + P(q_{12}|\theta H)$$

which put more weight on q'

Theory of Probability revisited
Significance tests: one new parameter
Improper priors for tests

Vague proper priors are not the solution

Taking a proper prior and take a "very large" variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Lindley's paradox

Example (Normal case)

If testing

$$H_0 : \theta = 0$$

when observing

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

under a normal $\mathcal{N}(0, \alpha)$ prior

$$B_{01}(x) \overset{\alpha \longrightarrow \infty}{\longrightarrow} 0$$

Often dubbed *Jeffreys–Lindley paradox...*

In terms of

$$t = \sqrt{n-1}\bar{x}/s', \qquad \nu = n-1$$

$$K \sim \sqrt{\frac{\pi\nu}{2}}\left(1+\frac{t^2}{\nu}\right)^{-1/2\nu+1/2}.$$

(...) The variation of K with t is much more important than the variation with ν (V, §5.2).

But ToP misses the point that under H_0 $t \sim \mathcal{T}_\nu$ so does not vary much with ν while ν goes to ∞...

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$B_{12} = \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!}e^{-\lambda\beta}\mathrm{d}\lambda}{\dfrac{1}{M}\displaystyle\sum_m \frac{x}{m-x+1}\,\frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha,\beta)$$

$$= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)}\beta^{-x}\Big/\frac{1}{M}\sum_m \frac{x}{m-x+1}$$

$$= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1}\beta^{-x}\Big/\frac{1}{M}\sum_m \frac{x}{m-x+1}$$

depends on choice of $\alpha(\beta)$ or $\beta(\alpha) \longrightarrow 0$

Learning from the sample

Definition (Learning sample)

Given an improper prior π, $(x_1, \ldots, x_n)$ is a *learning sample* if $\pi(\cdot|x_1, \ldots, x_n)$ is proper and a *minimal learning sample* if none of its subsamples is a learning sample

There is just enough information in a minimal learning sample to make inference about θ under the prior π

Pseudo-Bayes factors

Idea

Use one part $x_{[i]}$ of the data x to make the prior proper:

- π_i improper but $\pi_i(\cdot|x_{[i]})$ proper
- and
$$\frac{\int f_i(x_{[n/i]}|\theta_i)\ \pi_i(\theta_i|x_{[i]})\mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\ \pi_j(\theta_j|x_{[i]})\mathrm{d}\theta_j}$$
independent of normalizing constant
- Use remaining $x_{[n/i]}$ to run test as if $\pi_j(\theta_j|x_{[i]})$ is the true prior

Theory of Probability revisited
Significance tests: one new parameter
Pseudo-Bayes factors

Motivation

- Provides a working principle for improper priors
- Gather enough information from data to achieve properness
- and use this properness to run the test on remaining data
- does not use x twice as in Aitkin's (1991)

Theory of Probability revisited
Significance tests: one new parameter
Pseudo-Bayes factors

Details

Since $\pi_1(\theta_1|x_{[i]}) = \dfrac{\pi_1(\theta_1)f^1_{[i]}(x_{[i]}|\theta_1)}{\int \pi_1(\theta_1)f^1_{[i]}(x_{[i]}|\theta_1)\mathrm{d}\theta_1}$

$$\begin{aligned} B_{12}(x_{[n/i]}) &= \frac{\int f^1_{[n/i]}(x_{[n/i]}|\theta_1)\pi_1(\theta_1|x_{[i]})\mathrm{d}\theta_1}{\int f^2_{[n/i]}(x_{[n/i]}|\theta_2)\pi_2(\theta_2|x_{[i]})\mathrm{d}\theta_2} \\ &= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}\,\frac{\int \pi_2(\theta_2)f^2_{[i]}(x_{[i]}|\theta_2)\mathrm{d}\theta_2}{\int \pi_1(\theta_1)f^1_{[i]}(x_{[i]}|\theta_1)\mathrm{d}\theta_1} \\ &= B^N_{12}(x)B_{21}(x_{[i]}) \end{aligned}$$

© Independent of scaling factor!

Theory of Probability revisited
Significance tests: one new parameter
Pseudo-Bayes factors

Unexpected problems!

- depends on the choice of $x_{[i]}$
- many ways of combining pseudo-Bayes factors
  - AIBF $= B^N_{ji}\dfrac{1}{L}\sum_\ell B_{ij}(x_{[\ell]})$
  - MIBF $= B^N_{ji}$ med$[B_{ij}(x_{[\ell]})]$
  - GIBF $= B^N_{ji}$ exp $\dfrac{1}{L}\sum_\ell \log B_{ij}(x_{[\ell]})$
- not often an exact Bayes factor
- and thus lacking inner coherence

$$B_{12} \neq B_{10}B_{02} \qquad \text{and} \qquad B_{01} \neq 1/B_{10}\,.$$

[Berger & Pericchi, 1996]

Unexpec'd problems (cont'd)

Example (Mixtures)

There is no sample size that proper-ises improper priors, except if a training sample is allocated to *each* component

Reason If

$$x_1, \ldots, x_n \sim \sum_{i=1}^{k} p_i f(x|\theta_i)$$

and

$$\pi(\theta) = \prod_i \pi_i(\theta_i) \text{ with } \int \pi_i(\theta_i) \mathrm{d}\theta_i = +\infty\,,$$

the posterior is never defined, because

$$\Pr(\text{"no observation from } f(\cdot|\theta_i)\text{"}) = (1-p_i)^n$$

Intrinsic priors

There may exist a true prior that provides the same Bayes factor

Example (Normal mean)

Take $x \sim \mathcal{N}(\theta, 1)$ with either $\theta = 0$ ($\mathfrak{M}_1$) or $\theta \neq 0$ ($\mathfrak{M}_2$) and $\pi_2(\theta) = 1$.
Then

$$\begin{array}{lll} B_{21}^{AIBF} = B_{21}\frac{1}{\sqrt{2\pi}}\frac{1}{n}\sum_{i=1}^{n} e^{-x_1^2/2} & \approx B_{21} & \text{for } \mathcal{N}(0,2) \\ B_{21}^{MIBF} = B_{21}\frac{1}{\sqrt{2\pi}} e^{-\text{med}(x_1^2)/2} & \approx 0.93 B_{21} & \text{for } \mathcal{N}(0,1.2) \end{array}$$

[Berger and Pericchi, 1998]

When such a prior exists, it is called an **intrinsic prior**

Intrinsic priors (cont'd)

Example (Exponential scale)

Take $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \exp(\theta - x)\mathbb{I}_{x \geq \theta}$
and $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$, with $\pi_1(\theta) = 1$
Then

$$B_{10}^{A} = B_{10}(x)\frac{1}{n}\sum_{i=1}^{n}\left[e^{x_i - \theta_0} - 1\right]^{-1}$$

is the Bayes factor for

$$\pi_2(\theta) = e^{\theta_0 - \theta}\left\{1 - \log\left(1 - e^{\theta_0 - \theta}\right)\right\}$$

Most often, however, the pseudo-Bayes factors do not correspond to any true Bayes factor

[Berger and Pericchi, 2001]

Fractional Bayes factor

Idea

use directly the likelihood to separate training sample from testing sample

$$B_{12}^{F} = B_{12}(x)\frac{\int L_2^b(\theta_2)\pi_2(\theta_2)d\theta_2}{\int L_1^b(\theta_1)\pi_1(\theta_1)d\theta_1}$$

[O'Hagan, 1995]

Proportion b of the sample used to gain proper-ness

Fractional Bayes factor (cont'd)

Example (Normal mean)

$$B_{12}^F = \frac{1}{\sqrt{b}}\, e^{n(b-1)\bar{x}_n^2/2}$$

corresponds to exact Bayes factor for the prior $\mathcal{N}\left(0, \frac{1-b}{nb}\right)$

- If b constant, prior variance goes to 0
- If $b = \frac{1}{n}$, prior variance stabilises around 1
- If $b = n^{-\alpha}$, $\alpha < 1$, prior variance goes to 0 too.

Random effect models

In **ToP**, systematic errors (V, §5.6) correspond to random effect models

$$x_{ij} = y_{ij} + \epsilon_i$$

with $y_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \sigma^2$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$
Test of a systematic error is then equivalent to testing $\tau = 0$

But use of χ^2 test and MLE's....!

Comparison with classical tests

Standard answer

Definition (p-value)

The *p-value* $p(x)$ associated with a test is the largest significance level for which H_0 is rejected

Note

An alternative definition is that a p-value is distributed uniformly under the null hypothesis.

p-value

Example (Normal mean)

Since the UUMP test is $\{|x| > k\}$, standard p-value

$$\begin{aligned} p(x) &= \inf\{\alpha;\, |x| > k_\alpha\} \\ &= P^X(|X| > |x|), \qquad X \sim \mathcal{N}(0,1) \\ &= 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)]. \end{aligned}$$

Thus, if $x = 1.68$, $p(x) = 0.10$ and, if $x = 1.96$, $p(x) = 0.05$.

Problems with p-values

- Evaluation of the **wrong** quantity, namely the probability to exceed the observed quantity.(wrong conditionin)

  What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it had not predicted observable results that have not occurred (VII, §7.2)

- No transfer of the UMP optimality
- No decisional support (occurences of inadmissibility)
- Evaluation only under the null hypothesis
- Huge numerical difference with the Bayesian range of answers

Bayesian lower bounds

For illustration purposes, consider a class $\mathscr{G}$ of prior distributions

$$B(x,\mathscr{G}) = \inf_{g\in\mathscr{G}} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta)g(\theta)\,d\theta},$$
$$P(x,\mathscr{G}) = \inf_{g\in\mathscr{G}} \frac{f(x|\theta_0)}{f(x|\theta_0)+\int_\Theta f(x|\theta)g(\theta)\,d\theta}$$

when $\varrho_0 = 1/2$ or

$$B(x,\mathscr{G}) = \frac{f(x|\theta_0)}{\sup_{g\in\mathscr{G}}\int_\Theta f(x|\theta)g(\theta)d\theta}, \quad P(x,\mathscr{G}) = \left[1+\frac{1}{\underline{B}(x,\mathscr{G})}\right]^{-1}.$$

Resolution

Lemma

If there exists a mle for θ, $\hat\theta(x)$, the solutions to the Bayesian lower bounds are

$$B(x,\mathscr{G}) = \frac{f(x|\theta_0)}{f(x|\hat\theta(x))}, \quad \underline{P}P(x,\mathscr{G}) = \left[1+\frac{f(x|\hat\theta(x))}{f(x|\theta_0)}\right]^{-1}$$

respectively

Normal case

When $x\sim\mathscr{N}(\theta,1)$ and $H_0:\theta_0=0$, the lower bounds are

$$\underline{B}(x,G_A) = e^{-x^2/2} \quad \text{and} \quad \underline{P}(x,G_A) = \left(1+e^{x^2/2}\right)^{-1},$$

i.e.

p-value	0.10	0.05	0.01	0.001
P	0.205	0.128	0.035	0.004
B	0.256	0.146	0.036	0.004

[Quite different!]

Unilateral case

Different situation when $H_0 : \theta \leq 0$

- Single prior can be used both for H_0 and H_a
- Improper priors are therefore acceptable
- Similar numerical values compared with p-values

Unilateral agreement

Theorem

When $x \sim f(x-\theta)$, with f symmetric around 0 and endowed with the monotone likelihood ratio property, if $H_0 : \theta \leq 0$, the p-value $p(x)$ is equal to the lower bound of the posterior probabilities, $P(x, \mathscr{G}_{SU})$, when $\mathscr{G}_{SU}$ is the set of symmetric unimodal priors and when $x > 0$.

Reason:

$$p(x) = P_{\theta=0}(X > x) = \int_x^{+\infty} f(t)\,\mathrm{d}t = \inf_K \frac{1}{1 + \left[\frac{\int_{-K}^{0} f(x-\theta)\,\mathrm{d}\theta}{\int_{-K}^{K} f(x-\theta)}\,\mathrm{d}\theta\right]^{-1}}$$

Cauchy example

When $x \sim \mathscr{C}(\theta, 1)$ and $H_0 : \theta \leq 0$, lower bound inferior to p-value:

p-value	0.437	0.102	0.063	0.013	0.004
$\underline{P}$	0.429	0.077	0.044	0.007	0.002

Comments

- **ToP** very imprecise about choice of priors in the setting of tests
- **ToP** misses the difficulty of improper priors [coherent with earlier stance]
- but this problem still generates debates within the B community
- Some degree of goodness-of-fit testing but against fixed alternatives
- Persistence of the form

$$K \approx \sqrt{\frac{\pi n}{2}}\left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1/2}$$

but ν not so clearly defined...

Chapter 6: Significance tests: various complications

Theory of Probability revisited
Significance tests: various complications

Theory of Probability revisited
Significance tests: various complications
What's in there?!

Contents of Chapter 6

Certainly not a foundational chapter!!!

Some elementary remarks like

> *Anything that alters the prior probability of [the alternative parameter] will alter the rinferences about q' (VI, §6.0)*

and

> *One further possibility is that q and q' may not be initially equally probable (VI, §6.0)*

Theory of Probability revisited
Significance tests: various complications
What's in there?!

Repetition of the multiple alternatives

With several parameters under H_1, there are several embedded alternatives:

> *If the parameters are α, β, we can write q for the proposition $\alpha = \beta = 0$, q_α for $\alpha \neq 0, \beta = 0$, q_β for $\alpha =, \beta \neq 0$, and $q_{\alpha\beta}$ for $\alpha \neq 0, \beta \neq 0$ (VI, §6.1).*

Difficulty to order q_α and q_β reminiscent of Bernardo's reference priors but things get worse...

> *There is a best order of procedures, which is to assert the [hypothesis] that is most strongly supported, rejet those that are denied and proceed to consider further combinations (VI, §6.12)*

Re-enter Ockham explicitely!

Pluralitas non est ponenda sine neccesitate

Variation is random until the contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time, unless there is specific reason to the contrary. (...) This principle is workable and is a complete reversal of the usual notion of a 'principle of causality' (VI, §6.12)

Theory of Probability revisited
Significance tests: various complications
What's in there?!

Case of two location parameters (VI, §6.2)

ToP suggests to use a Cauchy prior $\mathcal{C}(0, \sigma^2)$ on the radius ρ and a uniform prior on the angle
Similar shape of the Bayes factor

$$K \approx \frac{n^{1/2}\pi}{2}\, t\, \left(\nu + t^2\right)^{-\nu/2+1}$$

Theory of Probability revisited
Significance tests: various complications
What's in there?!

Interesting extensions

Some hints at

- Hierarchical modelling (§6.3)

$$x_s \sim f(x - \alpha_\ell)\,, \quad \alpha_\ell \sim \tau g(\{\alpha_\ell - \alpha\}/\tau)$$

- Hidden Markov model (§6.4)

$$\mathbb{P} = \begin{bmatrix} \alpha + (1-\alpha)p_1 & (1-\alpha)p_2 & \cdots & (1-\alpha)p_r \\ (1-\alpha)p_1 & \alpha + (1-\alpha)p_2 & \cdots & (1-\alpha)p_r \\ & \cdots & & \\ (1-\alpha)p_1 & (1-\alpha)p_2 & \cdots & \alpha + (1-\alpha)p_r \end{bmatrix}$$

Theory of Probability revisited
Significance tests: various complications
What's in there?!

Un-interesting disgressions

Section §6.5 very windy about the nature of deduction and the approximation of point null hypotheses by interval representations

...by extending the meaning of q so as to say that the new parameter is not 0 but may be anywhere in some finite range. (...) I think, however, that it is both impossible and undesirable. (...) If there is anything to suggest a range of possible values it should go into the statement of q', not of q (VI, §6.5).

Model choice and model comparison

Choice of models

Several models available for the same observation

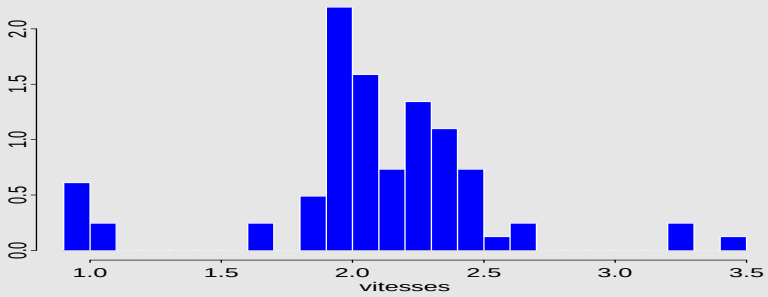
$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

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

Example (Galaxy normal mixture)

Set of observations of radial speeds of 82 galaxies possibly modelled as a mixture of normal distributions

$$\mathfrak{M}_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i}\mathcal{N}(\mu_{\ell i}, \sigma_{\ell i}^2)$$



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

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

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

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

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Ockham's rule)
  - how to fight overfitting for nested models

Which loss ?

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Warning

Parameters common to several models must be treated as separate entities!

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces
  - summation over many models (2^k)

Compatibility principle

Difficulty of finding simultaneously priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$
Easier to start from a single prior on a "big" model and to derive the others from a coherence principle

[Dawid & Lauritzen, 2000]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^\perp(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^\perp(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^\perp)) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Or we can look instead at the posterior distribution of

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^\perp))$$

[Goutis & Robert, 1998]

Theory of Probability revisited
Significance tests: various complications
Compatible priors

Operational principle for variable selection

Selection rule

Among all subsets $\mathcal{A}$ of covariates such that

$$d(\mathfrak{M}_g, \mathfrak{M}_\mathcal{A}) = \mathbb{E}_x[d(f_g(\cdot|x,\alpha), f_\mathcal{A}(\cdot|x_\mathcal{A}, \alpha^\perp))] < \epsilon$$

select the submodel with the smallest number of variables.

[Dupuis & Robert, 2001]

Theory of Probability revisited
Significance tests: various complications
Compatible priors

Kullback proximity

Alternative to above

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

$$\pi_2 = \arg\min_{\pi_2} \int \log\left(\frac{m_1(x;\pi_1)}{m_2(x;\pi_2)}\right) m_1(x;\pi_1)\,\mathrm{d}x$$

Theory of Probability revisited
Significance tests: various complications
Compatible priors

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...

Case of exponential families

Models

$$\mathfrak{M}_1 : \{f_1(x|\theta), \theta \in \Theta\}$$

and

$$\mathfrak{M}_2 : \{f_2(x|\lambda), \lambda \in \Lambda\}$$

sub-model of $\mathcal{M}_1$,

$$\forall \lambda \in \Lambda, \exists\, \theta(\lambda) \in \Theta, f_2(x|\lambda) = f_1(x|\theta(\lambda))$$

Both $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are natural exponential families

$$\begin{aligned} f_1(x|\theta) &= h_1(x)\exp(\theta^{\mathsf{T}} t_1(x) - M_1(\theta)) \\ f_2(x|\lambda) &= h_2(x)\exp(\lambda^{\mathsf{T}} t_2(x) - M_2(\lambda)) \end{aligned}$$

Conjugate priors

Parameterised (conjugate) priors

$$\begin{aligned} \pi_1(\theta; s_1, n_1) &= C_1(s_1, n_1)\exp(s_1^{\mathsf{T}}\theta - n_1 M_1(\theta)) \\ \pi_2(\lambda; s_2, n_2) &= C_2(s_2, n_2)\exp(s_2^{\mathsf{T}}\lambda - n_2 M_2(\lambda)) \end{aligned}$$

with closed form marginals $(i = 1, 2)$

$$m_i(x; s_i, n_i) = \int f_i(x|u)\pi_i(u)du = \frac{h_i(x)C_i(s_i, n_i)}{C_i(s_i + t_i(x), n_i + 1)}$$

Conjugate compatible priors

(Q.) Existence and unicity of Kullback-Leibler projection

$$\begin{aligned} (s_2^*, n_2^*) &= \arg\min_{(s_2,n_2)} \mathfrak{KL}(m_1(\cdot; s_1, n_1), m_2(\cdot; s_2, n_2)) \\ &= \arg\min_{(s_2,n_2)} \int \log\left(\frac{m_1(x; s_1, n_1)}{m_2(x; s_2, n_2)}\right) m_1(x; s_1, n_1)\mathrm{d}x \end{aligned}$$

A sufficient condition

Sufficient statistic $\psi = (\lambda, -M_2(\lambda))$

Theorem (Existence)

If, for all (s_2, n_2), the matrix

$$\mathbb{V}^{\pi_2}_{s_2,n_2}[\psi] - \mathbb{E}^{m_1}_{s_1,n_1}\left[\mathbb{V}^{\pi_2}_{s_2,n_2}(\psi|x)\right]$$

is semi-definite negative, the conjugate compatible prior exists, is unique and satisfies

$$\begin{aligned} \mathbb{E}^{\pi_2}_{s_2^*,n_2^*}[\lambda] - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(\lambda|x)] &= 0 \\ \mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)) - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)|x)] &= 0. \end{aligned}$$

An application to linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

$$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^{\mathsf{T}} X_1)^{-1}\right)$$

where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$

2. $\mathfrak{M}_2$:

$$y|\beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2), \quad \beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^{\mathsf{T}} X_2)^{-1}\right),$$

where X_2 is a $(n \times k_2)$ matrix with $\text{span}(X_2) \subseteq \text{span}(X_1)$

For a fixed (s_1, n_1), we need the projection $(s_2, n_2) = (s_1, n_1)^{\perp}$

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Theorem

Conditional on σ^2, the conjugate compatible prior of $\mathfrak{M}_2$ wrt $\mathfrak{M}_1$ is

$$\beta_2|X_2, \sigma^2 \sim \mathcal{N}\left(s_2^*, \sigma^2 n_2^* (X_2^T X_2)^{-1}\right)$$

with

$$\begin{aligned} s_2^* &= (X_2^T X_2)^{-1} X_2^T X_1 s_1 \\ n_2^* &= n_1 \end{aligned}$$

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0,1\}^p$ indicates inclusion/exclusion of variables by a binary representation, e.g. $\gamma = 101001011$ means that x_1, x_3, x_5, x_7 and x_8 are included.

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ indices of those variables and $t_0(\gamma)$ indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned} \beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\ X_{t_1(\gamma)} &= \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,q_\gamma}(\gamma)}\right]. \end{aligned}$$

Submodel $\mathfrak{M}_\gamma$ is thus

$$y|\beta, \gamma, \sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)}\beta_{t_1(\gamma)}, \sigma^2 I_n\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^{\mathsf{T}}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

▸ Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^{\mathsf{T}}X_{t_1(\gamma)}\right)^{-1}X_{t_1(\gamma)}^{\mathsf{T}}X\tilde{\beta}, c\sigma^2\left(X_{t_1(\gamma)}^{\mathsf{T}}X_{t_1(\gamma)}\right)^{-1}\right)$$

[Surprise!]

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model (and a priori independence between the presence/absence of variables)
Typically, when no prior information is available, $\tau_1 = \ldots = \tau_p = 1/2$, ie a uniform prior

$$\pi(\gamma) = 2^{-p}$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2}\left[y^{\mathsf{T}}y - \frac{cy^{\mathsf{T}}P_1y}{c+1} + \frac{\tilde{\beta}^{\mathsf{T}}X^{\mathsf{T}}P_1X\tilde{\beta}}{c+1} - \frac{2y^{\mathsf{T}}P_1X\tilde{\beta}}{c+1}\right]^{-n/2}.$$

Conditionally on γ, posterior distributions of β and σ^2:

$$\begin{aligned}\beta_{t_1(\gamma)}|\sigma^2, y, \gamma &\sim \mathcal{N}\left[\frac{c}{c+1}(U_1y + U_1X\tilde{\beta}/c), \frac{\sigma^2c}{c+1}\left(X_{t_1(\gamma)}^{\mathsf{T}}X_{t_1(\gamma)}\right)^{-1}\right],\\ \sigma^2|y, \gamma &\sim \mathcal{IG}\left[\frac{n}{2}, \frac{y^{\mathsf{T}}y}{2} - \frac{cy^{\mathsf{T}}P_1y}{2(c+1)} + \frac{\tilde{\beta}^{\mathsf{T}}X^{\mathsf{T}}P_1X\tilde{\beta}}{2(c+1)} - \frac{y^{\mathsf{T}}P_1X\tilde{\beta}}{c+1}\right].\end{aligned}$$

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)$$

▸ Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

Taking $\tilde{\beta} = 0_{p+1}$ and c large does not work

Influence of c

Erase influence

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i x_i + \sum_{i=1}^{3}\beta_{i+3}x_i^2 + \beta_7 x_1x_2 + \beta_8 x_1x3 + \beta_9 x_2x_3 + \beta_{10}x_1x_2x_3, \sigma^2 I_n\right)$$

where the x_is are iid $\mathcal{U}(0,10)$

[Casella & Moreno, 2004]

True model: two predictors x_1 and x_2, i.e. $\gamma^* = 110\ldots0$, $(\beta_0,\beta_1,\beta_2) = (5,1,3)$, and $\sigma^2 = 4$.

Theory of Probability revisited
Significance tests: various complications
Variable selection

Influence of c^2

$t_1(\gamma)$	$c = 10$	$c = 100$	$c = 10^3$	$c = 10^4$	$c = 10^6$
0,1,2	0.04062	0.35368	0.65858	0.85895	0.98222
0,1,2,7	0.01326	0.06142	0.08395	0.04434	0.00524
0,1,2,4	0.01299	0.05310	0.05805	0.02868	0.00336
0,2,4	0.02927	0.03962	0.00409	0.00246	0.00254
0,1,2,8	0.01240	0.03833	0.01100	0.00126	0.00126

Theory of Probability revisited
Significance tests: various complications
Variable selection

Noninformative case (cont'd)

In the noninformative setting,

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[y^{\mathsf{T}}y - \frac{c}{c+1}y^{\mathsf{T}}P_1 y\right]^{-n/2}$$

converges for all y's

Theory of Probability revisited
Significance tests: various complications
Variable selection

Casella & Moreno's example

$t_1(\gamma)$	$\sum_{i=1}^{10^6}\pi(\gamma\|y,c)\pi(c)$
0,1,2	0.78071
0,1,2,7	0.06201
0,1,2,4	0.04119
0,1,2,8	0.01676
0,1,2,5	0.01604

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.
Use of a Monte Carlo approximation of $\pi(\gamma|y)$

Gibbs sampling

- At $t = 0$, draw γ^0 from the uniform distribution on Γ
- At t, for $i = 1, \ldots, p$, draw
  $\gamma_i^t \sim \pi(\gamma_i|y, \gamma_1^t, \ldots, \gamma_{i-1}^t, \ldots, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})$

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.
True model with $n = 180$, $\sigma^2 = 4$ and seven predictor variables

$$x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20},$$
$$(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)$$

Gibbs approximation (cont'd)

Example (Simulated data (2))

γ	$\pi(\gamma\|y)$	$\widehat{\pi(\gamma\|y)}^{GIBBS}$
0,1,3,5,6,12,18,20	0.1893	0.1822
0,1,3,5,6,18,20	0.0588	0.0598
0,1,3,5,6,9,12,18,20	0.0223	0.0236
0,1,3,5,6,12,14,18,20	0.0220	0.0193
0,1,2,3,5,6,12,18,20	0.0216	0.0222
0,1,3,5,6,7,12,18,20	0.0212	0.0233
0,1,3,5,6,10,12,18,20	0.0199	0.0222
0,1,3,4,5,6,12,18,20	0.0197	0.0182
0,1,3,5,6,12,15,18,20	0.0196	0.0196

Gibbs ($T = 100,000$) results for $\tilde{\beta} = 0_{21}$ and $c = 100$

Processionary caterpillar

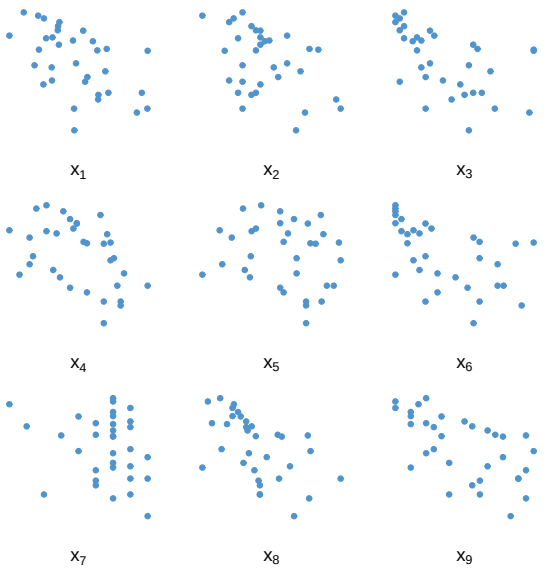
Influence of some forest settlement characteristics on the development of caterpillar colonies

Response y log-transform of the average number of nests of caterpillars per tree on an area of 500 square meters ($n = 33$ areas)

Processionary caterpillar (cont'd)

Potential explanatory variables

- x_1 altitude (in meters), x_2 slope (in degrees),
- x_3 number of pines in the square,
- x_4 height (in meters) of the tree at the center of the square,
- x_5 diameter of the tree at the center of the square,
- x_6 index of the settlement density,
- x_7 orientation of the square (from 1 if southb'd to 2 ow),
- x_8 height (in meters) of the dominant tree,
- x_9 number of vegetation strata,
- x_{10} mix settlement index (from 1 if not mixed to 2 if mixed).



Theory of Probability revisited
Significance tests: various complications
Variable selection

Bayesian regression output

	Estimate	BF	log10(BF)
(Intercept)	**9.2714**	**26.334**	**1.4205 (***)**
X1	**-0.0037**	**7.0839**	**0.8502 (**)**
X2	**-0.0454**	**3.6850**	**0.5664 (**)**
X3	**0.0573**	**0.4356**	**-0.3609**
X4	**-1.0905**	**2.8314**	**0.4520 (*)**
X5	**0.1953**	**2.5157**	**0.4007 (*)**
X6	**-0.3008**	**0.3621**	**-0.4412**
X7	**-0.2002**	**0.3627**	**-0.4404**
X8	**0.1526**	**0.4589**	**-0.3383**
X9	**-1.0835**	**0.9069**	**-0.0424**
X10	**-0.3651**	**0.4132**	**-0.3838**

evidence against H0: (****) decisive, (***) strong, (**) subtantial, (*) poor

Theory of Probability revisited
Significance tests: various complications
Variable selection

Bayesian variable selection

$t_1(\gamma)$	$\pi(\gamma\|y,X)$	$\widehat{\pi}(\gamma\|y,X)$
0,1,2,4,5	0.0929	0.0929
0,1,2,4,5,9	0.0325	0.0326
0,1,2,4,5,10	0.0295	0.0272
0,1,2,4,5,7	0.0231	0.0231
0,1,2,4,5,8	0.0228	0.0229
0,1,2,4,5,6	0.0228	0.0226
0,1,2,3,4,5	0.0224	0.0220
0,1,2,3,4,5,9	0.0167	0.0182
0,1,2,4,5,6,9	0.0167	0.0171
0,1,2,4,5,8,9	0.0137	0.0130

Noninformative G-prior model choice and Gibbs estimations

Postulate

Previous principle requires embedded models (or an encompassing model) and proper priors, while being hard to implement outside exponential families
Now we determine prior measures on two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, π_1 and π_2, directly by a compatibility principle.

Theory of Probability revisited
Significance tests: various complications
Symmetrised compatible priors

Generalised expected posterior priors

[Perez & Berger, 2000]

EPP Principle

Starting from reference priors π_1^N and π_2^N, substitute by prior distributions π_1 and π_2 that solve the system of integral equations

$$\pi_1(\theta_1) = \int_{\mathscr{X}} \pi_1^N(\theta_1 \,|\, x) m_2(x) \mathrm{d}x$$

and

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \,|\, x) m_1(x) \mathrm{d}x,$$

where x is an imaginary minimal training sample and m_1, m_2 are the marginals associated with π_1 and π_2 respectively.

Theory of Probability revisited
Significance tests: various complications
Symmetrised compatible priors

Motivations

- Eliminates the "imaginary observation" device and proper-isation through part of the data by integration under the "truth"
- Assumes that both models are *equally* valid and equipped with ideal unknown priors

$$\pi_i, \quad i = 1, 2,$$

that yield "true" marginals balancing each model wrt the other
- For a *given* π_1, π_2 is an **expected posterior prior**
Using both equations introduces symmetry into the game

Theory of Probability revisited
Significance tests: various complications
Symmetrised compatible priors

Dual properness

Theorem (Proper distributions)

If π_1 is a probability density then π_2 solution to

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \,|\, x) m_1(x) dx$$

is a probability density

© Both EPPs are either proper or improper

Bayesian coherence

Theorem (True Bayes factor)

If π_1 and π_2 are the EPPs and if their marginals are finite, then the corresponding Bayes factor

$$B_{1,2}(\mathbf{x})$$

is either a (true) Bayes factor or a limit of (true) Bayes factors.

Obviously only interesting when both π_1 and π_2 are improper.

Existence/Unicity

Theorem (Recurrence condition)

When both the observations and the parameters in both models are continuous, if the Markov chain with transition

$$Q\left(\theta_1' \,|\, \theta_1\right) = \int g\left(\theta_1, \theta_1', \theta_2, x, x'\right) \mathrm{d}x \mathrm{d}x' \mathrm{d}\theta_2$$

where

$$g\left(\theta_1, \theta_1', \theta_2, x, x'\right) = \pi_1^N\left(\theta_1' \,|\, x\right) f_2\left(x \,|\, \theta_2\right) \pi_2^N\left(\theta_2 \,|\, x'\right) f_1\left(x' \,|\, \theta_1\right),$$

is recurrent, *then there exists a solution to the integral equations, unique up to a multiplicative constant.*

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q\left(\theta_1' \,|\, \theta_1\right)$ has a dual transition density on Θ_2.
- There exists a parallel M chain on Θ_2 with identical properties; if one is (Harris) recurrent, so is the other.
- **Duality property** found both in the MCMC literature and in decision theory

  [Diebolt & Robert, 1992; Eaton, 1992]
- When Harris recurrence holds but the EPPs cannot be found, the Bayes factor can be approximated by MCMC simulation

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

$$\begin{aligned} \mathfrak{M}_1 &: f\left(x \,|\, \theta^*\right), \\ \mathfrak{M}_2 &: f\left(x \,|\, \theta\right), \theta \in \Theta. \end{aligned}$$

Default priors

$$\pi_1^N\left(\theta\right) = \delta_{\theta^*}\left(\theta\right) \text{ and } \pi_2^N\left(\theta\right) = \pi^N\left(\theta\right)$$

For x minimal training sample, consider the proper priors

$$\pi_1\left(\theta\right) = \delta_{\theta^*}\left(\theta\right) \text{ and } \pi_2\left(\theta\right) = \int \pi^N\left(\theta \,|\, x\right) f\left(x \,|\, \theta^*\right) \mathrm{d}x$$

Point null hypothesis testing (cont'd)

Then

$$\int \pi_1^N(\theta \mid x)\, m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) \int m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) = \pi_1(\theta)$$

and

$$\int \pi_2^N(\theta \mid x)\, m_1(x)\, \mathrm{d}x = \int \pi^N(\theta \mid x)\, f(x \mid \theta^*)\, \mathrm{d}x = \pi_2(\theta)$$

©$\pi_1(\theta)$ **and** $\pi_2(\theta)$ **are integral priors**

Note

Uniqueness of the Bayes factor
Integral priors and intrinsic priors coincide
[Moreno, Bertolino and Racugno, 1998]

Location models

Two location models

$$\begin{aligned} \mathfrak{M}_1 &: f_1(x \mid \theta_1) = f_1(x - \theta_1) \\ \mathfrak{M}_2 &: f_2(x \mid \theta_2) = f_2(x - \theta_2) \end{aligned}$$

Default priors

$$\pi_i^N(\theta_i) = c_i, \quad i = 1, 2$$

with minimal training sample size **one**
Marginal densities

$$m_i^N(x) = c_i, \quad i = 1, 2$$

Location models (cont'd)

In that case, $\pi_1^N(\theta_1)$ and $\pi_2^N(\theta_2)$ are integral priors **when** $c_1 = c_2$:

$$\begin{aligned} \int \pi_1^N(\theta_1 \mid x)\, m_2^N(x)\, \mathrm{d}x &= \int c_2 f_1(x - \theta_1)\, \mathrm{d}x = c_2 \\ \int \pi_2^N(\theta_2 \mid x)\, m_1^N(x)\, \mathrm{d}x &= \int c_1 f_2(x - \theta_2)\, \mathrm{d}x = c_1. \end{aligned}$$

© If the associated Markov chain is recurrent,

$$\pi_1^N(\theta_1) = \pi_2^N(\theta_2) = c$$

are the unique integral priors and they are intrinsic priors
[Cano, Kessler & Moreno, 2004]

Location models (cont'd)

Example (Normal versus double exponential)

$$\begin{aligned} \mathfrak{M}_1 &: \mathcal{N}(\theta, 1), \quad \pi_1^N(\theta) = c_1, \\ \mathfrak{M}_2 &: \mathcal{DE}(\lambda, 1), \quad \pi_2^N(\lambda) = c_2. \end{aligned}$$

Minimal training sample size one and posterior densities

$$\pi_1^N(\theta \mid x) = \mathcal{N}(x, 1) \text{ and } \pi_2^N(\lambda \mid x) = \mathcal{DE}(x, 1)$$

Location models (cont'd)

Example (Normal versus double exponential (2))

Transition $\theta \to \theta'$ of the Markov chain made of steps :

1. $x' = \theta + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0,1)$
2. $\lambda = x' + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{DE}(0,1)$
3. $x = \lambda + \varepsilon_3$, $\varepsilon_3 \sim \mathcal{DE}(0,1)$
4. $\theta' = x + \varepsilon_4$, $\varepsilon_4 \sim \mathcal{N}(0,1)$

$$\text{i.e.} \qquad \theta' = \theta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$$

random walk in θ with finite second moment, null recurrent
© **Resulting Lebesgue measures $\pi_1(\theta) = 1 = \pi_2(\lambda)$ invariant and unique solutions to integral equations**

Chapter 7: Frequency definitions and direct methods

In a dubious battle...

First, discussion against mathematical formalism that tries to build on intuition for finite state spaces

> *For continuous distributions there are an infinite number of possible cases, and the definition makes the probability, in the face of it, the ratio of two infinite numbers and therefore meaningless. (...) On the infinite population definition, any finite probability is the ratio of two infinite numbers and therefore is indeterminate (VII, §7.0)*

Not worth much except as an historical perspective

Dual representation

Next, discussion of dual meaning of Student's T distribution:

$$P(dz|x,\sigma,H) \propto (1+z^2)^{-1/2n}dz \qquad (1)$$

where (...)

$$z = \frac{x-\bar{x}}{s}\,.$$

My result is

$$P(dz|\bar{x},s,H) \propto (1+z^2)^{-1/2n}dz \qquad (4)$$

This is not the same thing as (1) since the data is different.

Explanation

While (1) is the (sampling) distribution of z as a transform of the data $(\bar{x},s)$, (4) is the (posterior) distribution of the mean parameter x given the data.
Instance of a (Fisherian) pivotal quantity

Warnin! Dependence on the prior distribution

there is only one distrbution of the prior probability that can lead to it, namely

$$P(dxd\sigma|H) \propto dxd\sigma/\sigma$$

Criticism of frequentist tests

Rejection of Student's t test:

...we should reject a suggested value of x by such rule as this, but applying this in practice would imply that if x was known to be always the same we must accept it in 95 per cent. and reject it in 5 per cent. of the cases which hardly seems a satisfactory state of affairs. There is no positive virtue in rejecting a hypothesis in 5 per cent. of the cases where it is true, though it may be inevitable if we are to have any rule at all for rejecting it when it is false, that we shall sometimes reject it when it is true. In practice nobody would use the rule in this way if x was always the same; samples would always be combined (VII, §7.1).

Missing [degree of] freedom

Same criticism of Pearson's χ^2 test [if acceptance of Pearson's χ^2 estimation method...]

if there were n groups of observations [and if] m parameters had been found from the data, [Pearson] would form the integral (VII, §7.2)

$$P(\chi^2) = \int_\chi^\infty \chi^{n-m-1}e^{-1/2\chi}\,d\chi \Big/ \int_0^\infty \chi^{n-m-1}e^{-1/2\chi}\,d\chi$$

Should be $n-m-2$ to correspond to the standard χ^2_{n-m-1} approximation...

Criticism of p-values

One of **ToP** most quoted sentences:

> *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it had not predicted observable results that have not occurred (VII, §7.2)*

Even more to the point:

> *If P is small that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of P? (VII, §7.2)*

Criticism of p-values (cont'd)

Jeffreys defends the use of likelihood ratios [or inverse probability] versus p values (VII, §7.2)

> *...if the actual value is unknown the value of the power function is also unknown (...) [and] if we must choose between two definitely stated alternatives we should naturally take the one that gives the larger likelihood (VII, §7.5)*

Criticism of p-values (cont'd)

Acceptance of posterior probability statements related to confidence assessments:

> *...several of the P integrals have a definitive place in the present theory, in problems of pure estimation. For the normal law with a known standard error, the total are of the tail represents the probability, given the data, that the estimated difference has the wrong sign (VII, §7.21)*

As for instance in design:

> *...the P integral found from the difference between the mean yields of two varieties gives correctly the probability on the data that the estimates are in the wrong order, which is what is required (VII, §7.21)*

Criticism of p-values (cont'd)

But does not make sense for testing point null hypotheses

> *If some special value has to be excluded before we can assert any other value, what is the best rule on the data available for deciding to retain it or adopt a new one? (VII, §7.21)*

And **ToP** finds no justification in the .05 golden rule

> *In itself it is fallacious [and] there is not the slightest reason to suppose that it gives the best standard (VII, §7.21)*

Another fundamental issue

Why are p values so bad?
Because they do not account for the alternative:

> *Is it of the slightest use to reject an hypothesis unless we have some idea of what to put in its place? (VII, §7.22)*

...and for the consequences of rejecting the null:

> *The test required, in fact, is not whether the null hypothesis is altogether satisfactory, but whether any suggested alternative is likely to give an improvement in representing future data (VII, §7.22)*

The disagreement with Fisher

Main points of contention (VII, §7.4)

...general agreement between Fisher and myself...

- *...hypothetical infinite population...*
- lack of conditioning
- *...use of the P integrals...*

Oooops!

> *...at that time, to my regret, I had not read 'Student's' papers and it was not till considerably later that I saw the intimate relation between [Fisher's] methods and mine.*

Overall risk

Criticism of power as parameter dependent ◂ Power back
Use of average risk

> *...the expectation of the total fraction of mistakes will be*

$$2\int_{a_c}^{\infty} P(qda|H) + 2\int_{0}^{a_c}\int P(q'd\alpha da|H)\,.$$

> *Hence the total number of mistakes will be made a minimum if the line is drawn at the critical value that makes $K = 1$ (VII, §7.4).*

But bound becomes data-dependent!

Chapter 8: General questions

8 General questions

- Introduction
- Subjective prior
- Jeffrey's prior
- Missing alternatives
- Marginaliae
- Conclusion

Priors are not frequencies

First part (§8.0) focussing on the concept of prior distribution and the differences with a frequency based probability

> *The essence of the present theory is that no probability, direct, prior, or posterior is simply a frequency (VIII, §8.0).*

Extends this perspective to sampling distributions too [with hairy arguments!].

Common criticism

Next, discussion of the subjective nature of priors

> *Critics (...) usually say that the prior probability is 'subjective' (...) or refer to the vagueness of previous knowledge as an indication that the prior probability cannot be assessed (VIII, §8.0).*

Conditional features of probabilities

Long argument about the subjective nature of knowledge

> *What the present theory does is to resolve the problem by making a sharp distinction between general principles, which are deliberately designed to say nothing about what experience is possible, and, on the other hand, propositions that do concern experience and are in the first place merely considered among alternatives (VIII, §8.1).*

and definition of probability

> *The probability of a proposition irrespective of the data has no meaning and is simply an unattainable ideal (VIII, §8.1).*

Noninformative priors

ToP then advances the use of Jeffreys' priors as the answer to missing prior information

> *A prior probability used to express ignorance is* merely *the formal statement of ignorance (VIII, §8.1).*

Overlooks the lack of uniqueness of such priors

Missing alternatives

Next section §8.2 fairly interesting in that **ToP** discusses the effect of a missing alternative

> *We can never rule out the possibility that some new explanantion may be suggested of any set of experimental facts (VIII, §8.2).*

Seems partly wrong though...

Missing alternatives (cont'd)

Indeed, if H_0 tested against H_1, Bayes factor is

$$B_{01}^{\pi} = \frac{\int f_0(x|\theta_0)\pi_0(\mathrm{d}\theta_0)}{\int f_1(x|\theta_1)\pi_1(\mathrm{d}\theta_1)}$$

while if another (exclusive) alernative H_2 is introduced, it would be

$$B_{01}^{\pi} = \frac{\int f_0(x|\theta_0)\pi_0(\mathrm{d}\theta_0)}{\omega_1 \int f_1(x|\theta_1)\pi_1(\mathrm{d}\theta_1) + (1-\omega_1)\int f_2(x|\theta_2)\pi_2(\mathrm{d}\theta_2)}$$

where ω_1 relative prior weight of H_1 vs H_2
Basically biased in favour of H_0

Marginaliae

The remaining sections are not very interesting from a Bayesian point of view [but may be so from an epistemiological point of view (quantum theory, relativity, "rejection of unobservables", realism vs. idealism)...]

The end is near!!!

Conclusive section about **ToP** principles

> *...we have first the main principle that the ordinary common-sense notion of probability is capable of consistent treatment (VIII, §8.6).*

...although consistency is not precisely defined.

> *The principle of inverse probability is a theorem (VIII, §8.6).*

Theory of Probability revisited
General questions
Conclusion

Jeffreys' priors at the center of this theory:

> *The prior probabilities needed to express ignorance of the value of a quantity to be estimated, where there is nothing to call special attention to a particular value are given by an invariance theory (VIII, §8.6).*

with adequate changes for testing hypotheses:

> *Where a question of significance arises, that is, where previous considerations call attention to some particular value, half, or possibly some smaller fraction, of the prior probability is concentrated at that value (VIII, §8.6).*

Theory of Probability revisited
General questions
Conclusion

Main results

1. *a proof independent of limiting processes that the whole information is contained in the likelihood*
2. *a development of pure estimation processes without further hypothesis*
3. *a general theory of significance tests*
4. *an account of how in certain conditions a law can reach a high probability*

Theory of Probability revisited
General questions
Conclusion

Corresponing remaining problems in **ToP**

1. information also contained in prior distribution
2. choice of estimation procedure never explicited
3. complete occultation of the infinite mass problem
4. no true theory of goodness of fit tests

but...

> *...it is enough (VIII, §8.8).*