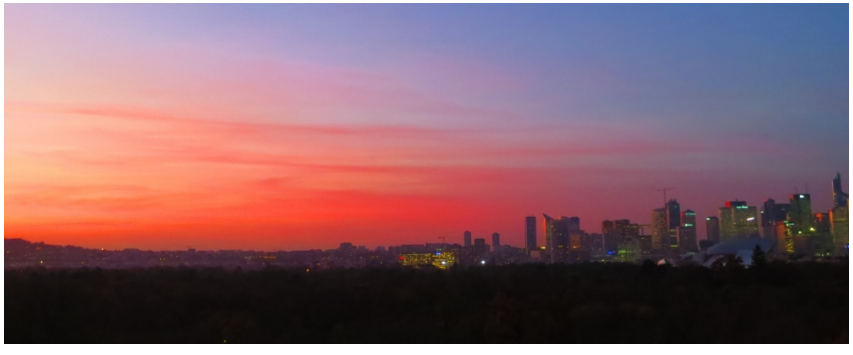
Bayesian tests of hypotheses

Christian P. Robert
Université Paris-Dauphine, Paris & University of Warwick, Coventry



Joint work with K. Kamary, K. Mengersen & J. Rousseau

Outline

Bayesian testing of hypotheses

Noninformative solutions

Testing via mixtures

Paradigm shift

Testing issues

Hypothesis testing

- central problem of statistical inference
- witness the recent ASA's statement on p-values (Wasserstein, 2016)
- dramatically differentiating feature between classical and Bayesian paradigms
- wide open to controversy and divergent opinions, includ. within the Bayesian community
- non-informative Bayesian testing case mostly unresolved, witness the Jeffreys–Lindley paradox

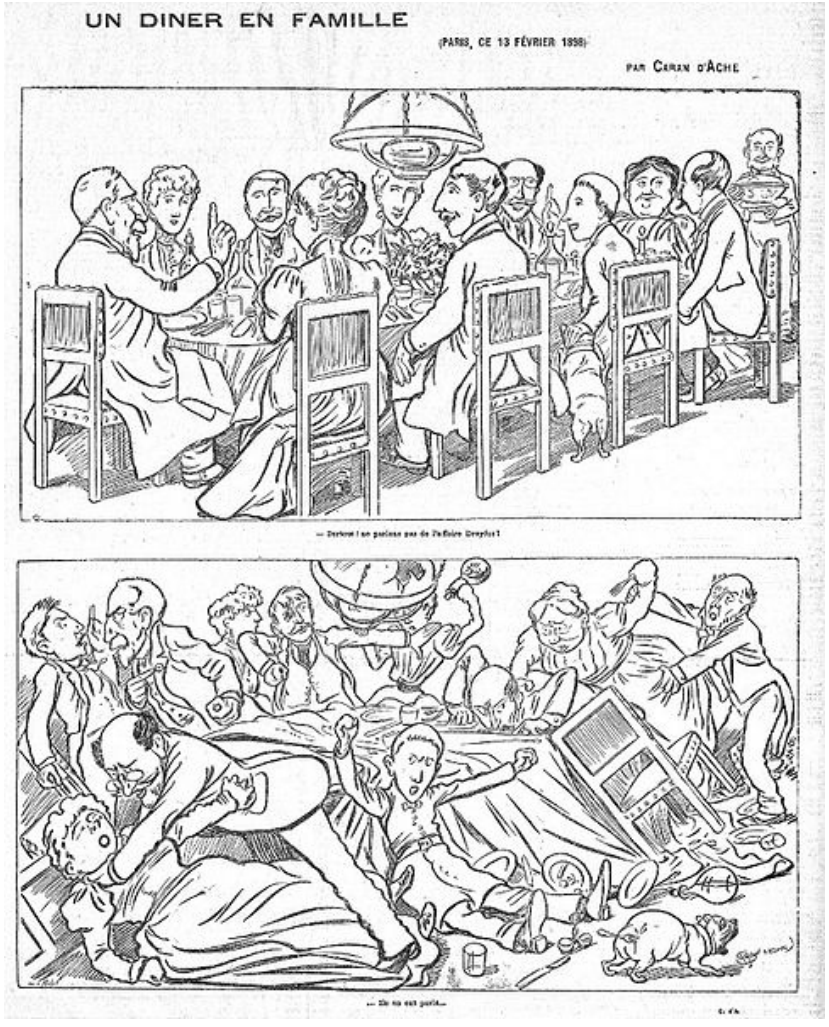
[Berger (2003), Mayo & Cox (2006), Gelman (2008)]

"proper use and interpretation of the *p*-value"

"Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

"By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

[Wasserstein, 2016]

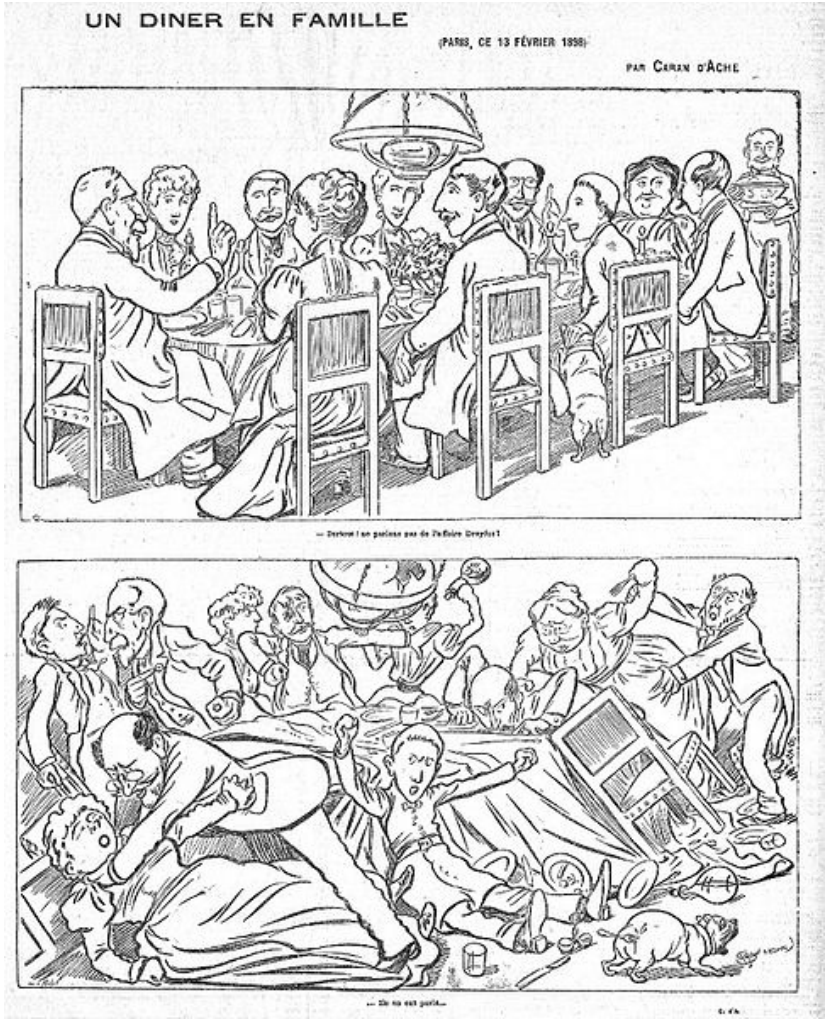


"proper use and interpretation of the *p*-value"

"Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

"By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

[Wasserstein, 2016]



Bayesian testing of hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data "best"
- mostly accepted perspective: it does not primarily seek to identify which model is "true", but compares fits
- tools like Bayes factor naturally include a penalisation addressing model complexity, mimicked by Bayes Information (BIC) and Deviance Information (DIC) criteria
- posterior predictive tools successfully advocated in Gelman et al. (2013) even though they involve double use of data

Bayesian testing of hypotheses

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data "best"
- mostly accepted perspective: it does not primarily seek to identify which model is "true", but compares fits
- tools like Bayes factor naturally include a penalisation addressing model complexity, mimicked by Bayes Information (BIC) and Deviance Information (DIC) criteria
- posterior predictive tools successfully advocated in Gelman et al. (2013) even though they involve double use of data

Bayesian tests 101

Associated with the risk

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

Bayes test

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

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

Bayesian tests 101

Associated with the risk

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

Bayes test

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

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

Bayesian tests 102

Weights errors differently under both hypotheses:

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

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

Bayesian tests 102

Weights errors differently under both hypotheses:

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

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

A function of posterior probabilities

Definition (Bayes factors)

For hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$

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

[Jeffreys, **ToP**, 1939, V, §5.01]

Bayes rule under $0-1$ loss: acceptance if

$$\mathfrak{B}_{01} > \{(1-\pi(\Theta_0))/a_1\}/\{\pi(\Theta_0)/a_0\}$$

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:
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ above 2, evidence *decisive*

[...fairly arbitrary!]

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:
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ above 2, evidence *decisive*

[...fairly arbitrary!]

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:
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ above 2, evidence *decisive*

[...fairly arbitrary!]

self-contained concept

Outside decision-theoretic environment:

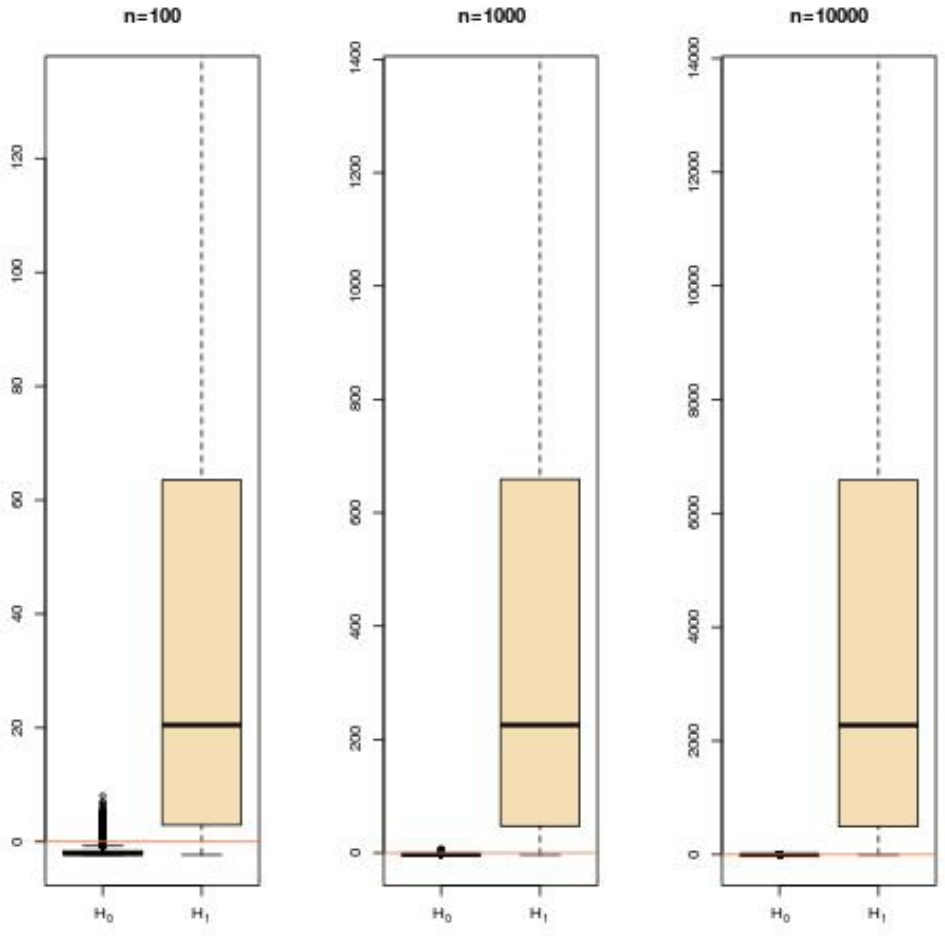
- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio
- Jeffreys' scale of evidence:
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(\mathfrak{B}_{10}^{\pi})$ above 2, evidence *decisive*

[...fairly arbitrary!]

consistency

Example of a normal $\bar{X}_n \sim \mathcal{N}(\mu, 1/n)$ when $\mu \sim \mathcal{N}(0, 1)$, leading to

$$\mathfrak{B}_{01} = (1+n)^{-1/2} \exp\{n^2 \bar{x}_n^2 / 2(1+n)\}$$



Some difficulties

- tension between using (i) posterior probabilities justified by binary loss function but depending on unnatural prior weights, and (ii) Bayes factors that eliminate dependence but escape direct connection with posterior, unless prior weights are integrated within loss
- delicate interpretation (or calibration) of strength of the Bayes factor towards supporting a given hypothesis or model, because *not* a Bayesian decision rule
- similar difficulty with posterior probabilities, with tendency to interpret them as p-values: only report of respective strengths of fitting data to both models

Some difficulties

- tension between using (i) posterior probabilities justified by binary loss function but depending on unnatural prior weights, and (ii) Bayes factors that eliminate dependence but escape direct connection with posterior, unless prior weights are integrated within loss
- referring to a fixed and arbitrary cuttoff value falls into the same difficulties as regular p-values
- no "third way" like opting out from a decision

Some further difficulties

- long-lasting impact of prior modeling, i.e., choice of prior distributions on parameters of both models, despite overall consistency proof for Bayes factor
- discontinuity in valid use of improper priors since they are not justified in most testing situations, leading to many alternative and *ad hoc* solutions, where data is either used twice or split in artificial ways [or further tortured into confession]
- binary (*accept* vs. *reject*) outcome more suited for immediate decision (if any) than for model evaluation, in connection with rudimentary loss function [atavistic remain of Neyman-Pearson formalism]

Some additional difficulties

- related impossibility to ascertain simultaneous misfit or to detect outliers
- no assessment of uncertainty associated with decision itself besides posterior probability
- difficult computation of marginal likelihoods in most settings with further controversies about which algorithm to adopt
- strong dependence of posterior probabilities on conditioning statistics (ABC), which undermines their validity for model assessment
- temptation to create pseudo-frequentist equivalents such as q-values with even less Bayesian justifications
- © time for a paradigm shift
- back to some solutions

Historical appearance of Bayesian tests

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

...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

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

(Jeffreys, **ToP***, 1939, V, §5.0)*

A major refurbishment

> *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 $K[= \mathfrak{B}_{10}]$ would always be infinite (Jeffreys,* **ToP**, *V, §5.02)*

When the null hypothesis is supported by a set of measure 0 against Lebesgue measure, $\pi(\Theta_0) = 0$ for an absolutely continuous prior distribution

[End of the story?!]

A major refurbishment

When the null hypothesis is supported by a set of measure 0 against Lebesgue measure, $\pi(\Theta_0) = 0$ for an absolutely continuous prior distribution

[End of the story?!]

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)

A major refurbishment

When the null hypothesis is supported by a set of measure 0 against Lebesgue measure, $\pi(\Theta_0) = 0$ for an absolutely continuous prior distribution

[End of the story?!]

Using the prior probabilities $\pi(\Theta_0) = \rho_0$ and $\pi(\Theta_1) = \rho_1$,

$$\pi(\theta) = \rho_0\pi_0(\theta) + \rho_1\pi_1(\theta).$$

Point null hypotheses

"Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place?" H. Jeffreys, **ToP** *(p.390)*

Particular case $H_0 : \theta = \theta_0$
Take $\rho_0 = \Pr^\pi(\theta = \theta_0)$ and g_1 prior density under H_0^c.
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_0^c

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

and

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

Point null hypotheses

"Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place?" H. Jeffreys, **ToP** *(p.390)*

Particular case $H_0 : \theta = \theta_0$
Take $\rho_0 = \Pr^{\pi}(\theta = \theta_0)$ and g_1 prior density under H_0^c.
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_0^c

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

and

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

Noninformative proposals

Bayesian testing of hypotheses

Noninformative solutions

Testing via mixtures

Paradigm shift

what's special about the Bayes factor?!

- "The priors do not represent substantive knowledge of the parameters within the model
- Using Bayes' theorem, these priors can then be updated to posteriors conditioned on the data that were actually observed
- In general, the fact that different priors result in different Bayes factors should not come as a surprise
- The Bayes factor (...) balances the tension between parsimony and goodness of fit, (...) against overfitting the data
- In induction there is no harm in being occasionally wrong; it is inevitable that we shall be"

[Jeffreys, 1939; Ly et al., 2015]

what's wrong with the Bayes factor?!

- $(1/2, 1/2)$ partition between hypotheses has very little to suggest in terms of extensions
- central difficulty stands with issue of picking a prior probability of a model
- unfortunate impossibility of using improper priors in most settings
- Bayes factors lack direct scaling associated with posterior probability and loss function
- twofold dependence on subjective prior measure, first in prior weights of models and second in lasting impact of prior modelling on the parameters
- Bayes factor offers no window into uncertainty associated with decision
- further reasons in the ◂ summary

[Robert, 2016]

Lindley's paradox

In a normal mean testing problem,

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n)\,, \quad H_0 : \theta = \theta_0\,,$$

under Jeffreys prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, the Bayes factor

$$\mathfrak{B}_{01}(t_n) = (1+n)^{1/2} \exp\left(-nt_n^2/2[1+n]\right)\,,$$

where $t_n = \sqrt{n}|\bar{x}_n - \theta_0|/\sigma$, satisfies

$$\mathfrak{B}_{01}(t_n) \overset{n\longrightarrow\infty}{\longrightarrow} \infty$$

[assuming a fixed t_n]

[Lindley, 1957]

A strong impropriety

Improper priors not allowed in Bayes factors:

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 π_1 or π_2 cannot be coherently normalised while the normalisation matters in the Bayes factor $\mathfrak{B}_{12}$

Lack of mathematical justification for "common nuisance parameter" [and prior of]

[Berger, Pericchi, and Varshavsky, 1998; Marin and Robert, 2013]

A strong impropriety

Improper priors not allowed in Bayes factors:

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 π_1 or π_2 cannot be coherently normalised while the normalisation matters in the Bayes factor $\mathfrak{B}_{12}$

Lack of mathematical justification for "common nuisance parameter" [and prior of]

[Berger, Pericchi, and Varshavsky, 1998; Marin and Robert, 2013]

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc, which lacks complete proper Bayesian justification
  [Berger & Pericchi, 2001]

- use of *identical* improper priors on nuisance parameters,
- calibration via the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters, a notion already entertained by Jeffreys

[Berger et al., 1998; Marin & Robert, 2013]

- calibration via the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- *Péché de jeunesse:* equating the values of the prior densities at the point-null value θ_0,

$$\rho_0 = (1 - \rho_0)\pi_1(\theta_0)$$

[Robert, 1993]

- calibration via the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- calibration via the posterior predictive distribution, which uses the data twice
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- calibration via the posterior predictive distribution,
- matching priors, whose sole purpose is to bring frequentist and Bayesian coverages as close as possible

  [Datta & Mukerjee, 2004]
- use of score functions extending the log score function
- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- calibration via the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function

$$\log \mathfrak{B}_{12}(x) = \log m_1(x) - \log m_2(x) = S_0(x, m_1) - S_0(x, m_2)\,,$$

that are independent of the normalising constant
[Dawid et al., 2013; Dawid & Musio, 2015]

- non-local priors correcting default priors

On some resolutions of the paradox

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- calibration via the posterior predictive distribution,
- matching priors,
- use of score functions extending the log score function
- non-local priors correcting default priors towards more balanced error rates

[Johnson & Rossell, 2010; Consonni et al., 2013]

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

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

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

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)

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)

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)

Issues

- 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]

Issues

- depends on the choice of $x_{[i]}$
- many ways of combining pseudo-Bayes factors
  - AIBF $= B_{ji}^N \frac{1}{L} \sum_{\ell} B_{ij}(x_{[\ell]})$
  - MIBF $= B_{ji}^N \, \text{med}[B_{ij}(x_{[\ell]})]$
  - GIBF $= B_{ji}^N \, \exp \frac{1}{L} \sum_{\ell} \log B_{ij}(x_{[\ell]})$
- 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]

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

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.

© Call to external principles to pick the order of b

Bayesian predictive

"If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance." (BDA, p.143)

Use of posterior predictive

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)\pi(\theta|y)\,\mathrm{d}\theta$$

and measure of discrepancy $T(\cdot,\cdot)$
Replacing p-value

$$p(y|\theta) = \mathbb{P}(T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|\theta)$$

with Bayesian posterior p-value

$$\mathbb{P}(T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|y) = \int p(y|\theta)\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

Bayesian predictive

"If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance." (BDA, p.143)

Use of posterior predictive

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)\pi(\theta|y)\,\mathrm{d}\theta$$

and measure of discrepancy $T(\cdot,\cdot)$
Replacing p-value

$$p(y|\theta) = \mathbb{P}(\,T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|\theta)$$

with Bayesian posterior p-value

$$\mathbb{P}(\,T(y^{\text{rep}},\theta) \geqslant T(y,\theta)|y) = \int p(y|\theta)\pi(\theta|\mathbf{x})\,\mathrm{d}\theta$$

Issues

"the posterior predictive p-value is such a [Bayesian] probability statement, conditional on the model and data, about what might be expected in future replications. (BDA, p.151)

- sounds too much like a p-value...!
- relies on choice of $T(\cdot, \cdot)$
- seems to favour overfitting
- (again) using the data twice (once for the posterior and twice in the p-value)
- needs to be calibrated (back to 0.05?)
- general difficulty in interpreting
- where is the penalty for model complexity?

Changing the testing perspective

Bayesian testing of hypotheses

Noninformative solutions

Testing via mixtures

Paradigm shift

Paradigm shift

New proposal for a paradigm shift (!) in the Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- more extended use of improper priors

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 *or* 1

Paradigm shift

New proposal for a paradigm shift (!) in the Bayesian processing of hypothesis testing and of model selection

- convergent and naturally interpretable solution
- more extended use of improper priors

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 *or* 1

Paradigm shift

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 *or* 1

- Approach inspired from consistency result of Rousseau and Mengersen (2011) on estimated overfitting mixtures
- Mixture representation not directly equivalent to the use of a posterior probability
- Potential of a better approach to testing, while not expanding number of parameters
- Calibration of posterior distribution of the weight of a model, moving from artificial notion of posterior probability of a model

Encompassing mixture model

Idea: Given two statistical models,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1) \,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2) \,, \; \theta_2 \in \Theta_2 \,,$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2) \,, \; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

Note: Both models correspond to special cases of (1), one for $\alpha = 1$ and one for $\alpha = 0$

Draw inference on mixture representation (1), as if each observation was individually and independently produced by the mixture model

Encompassing mixture model

Idea: Given two statistical models,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1) \,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2) \,, \; \theta_2 \in \Theta_2 \,,$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2) \,, \; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

Note: Both models correspond to special cases of (1), one for $\alpha = 1$ and one for $\alpha = 0$

Draw inference on mixture representation (1), as if each observation was individually and independently produced by the mixture model

Encompassing mixture model

Idea: Given two statistical models,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1) \,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2) \,, \; \theta_2 \in \Theta_2 \,,$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2) \,, \; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

Note: Both models correspond to special cases of (1), one for $\alpha = 1$ and one for $\alpha = 0$

Draw inference on mixture representation (1), as if each observation was individually and independently produced by the mixture model

Inferential motivations

Sounds like approximation to the real model, but several definitive advantages to this paradigm shift:

- Bayes estimate of the weight α replaces posterior probability of model $\mathfrak{M}_1$, equally convergent indicator of which model is "true", while avoiding artificial prior probabilities on model indices, ω_1 and ω_2
- interpretation of estimator of α at least as natural as handling the posterior probability, while avoiding zero-one loss setting
- α and its posterior distribution provide measure of proximity to the models, while being interpretable as data propensity to stand within one model
- further allows for alternative perspectives on testing and model choice, like predictive tools, cross-validation, and information indices like WAIC

Computational motivations

- avoids highly problematic computations of the marginal likelihoods, since standard algorithms are available for Bayesian mixture estimation
- straightforward extension to a finite collection of models, with a larger number of components, which considers all models at once and eliminates least likely models by simulation
- eliminates difficulty of label switching that plagues both Bayesian estimation and Bayesian computation, since components are no longer exchangeable
- posterior distribution of α evaluates more thoroughly strength of support for a given model than the single figure outcome of a posterior probability
- variability of posterior distribution on α allows for a more thorough assessment of the strength of this support

Noninformative motivations

- additional feature missing from traditional Bayesian answers: a mixture model acknowledges possibility that, for a finite dataset, *both* models or *none* could be acceptable
- standard (proper and informative) prior modeling can be reproduced in this setting, but non-informative (improper) priors also are manageable therein, provided both models first reparameterised towards shared parameters, e.g. location and scale parameters
- in special case when all parameters are common

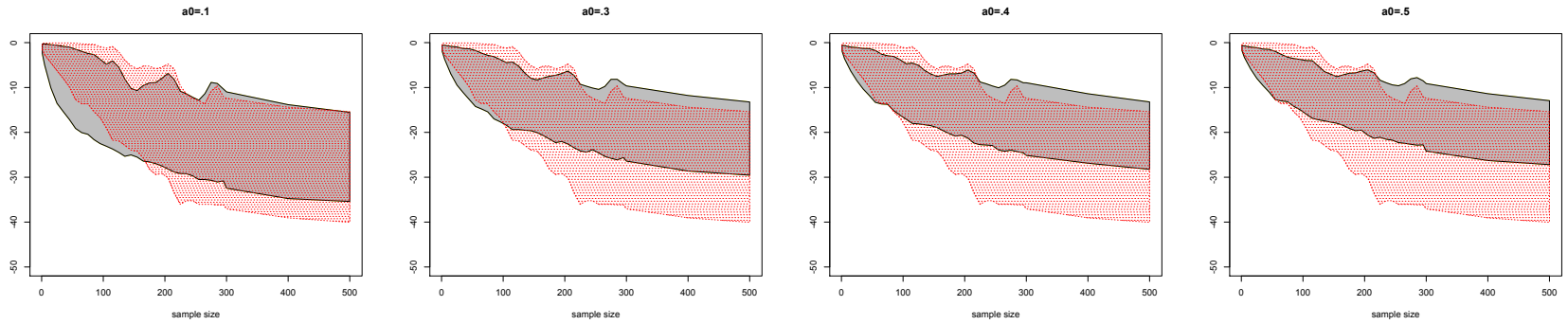
$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta) + (1-\alpha) f_2(x|\theta)\,, 0 \leqslant \alpha \leqslant 1$$

if θ is a location parameter, a flat prior $\pi(\theta) \propto 1$ is available

Weakly informative motivations

- using the *same* parameters or some *identical* parameters on both components highlights that opposition between the two components is not an issue of enjoying different parameters
- those common parameters are nuisance parameters, to be integrated out *[unlike Lindley's paradox]*
- prior model weights ω_i rarely discussed in classical Bayesian approach, even though linear impact on posterior probabilities. Here, prior modeling only involves selecting a prior on α, e.g., $\alpha \sim \mathcal{B}(a_0, a_0)$
- while a_0 impacts posterior on α, it always leads to mass accumulation near 1 or 0, i.e. favours most likely model
- sensitivity analysis straightforward to carry
- approach easily calibrated by parametric boostrap providing reference posterior of α under each model
- natural Metropolis–Hastings alternative

Comparison with posterior probability



Plots of ranges of $\log(n)$ $\log(1-\mathbb{E}[\alpha|x])$ *(gray color)* and $\log(1-p(\mathfrak{M}_1|x))$ *(red dotted)* over 100 $\mathcal{N}(0,1)$ samples as sample size n grows from 1 to 500. and α is the weight of $\mathcal{N}(0,1)$ in the mixture model. The shaded areas indicate the range of the estimations and each plot is based on a Beta prior with $a_0 = .1, .2, .3, .4, .5, 1$ and each posterior approximation is based on 10^4 iterations.

Towards which decision?

And if we have to make a decision?

soft consider behaviour of posterior under prior predictives

- or posterior predictive [e.g., prior predictive does not exist]
- boostrapping behaviour
- comparison with Bayesian non-parametric solution

hard rethink the loss function

Conclusion

- many applications of the Bayesian paradigm concentrate on the comparison of scientific theories and on testing of null hypotheses
- natural tendency to default to Bayes factors
- poorly understood sensitivity to prior modeling and posterior calibration

© Time is ripe for a paradigm shift

Down with Bayes factors!

© Time is ripe for a paradigm shift

- original testing problem replaced with a better controlled estimation target
- allow for posterior variability over the component frequency as opposed to deterministic Bayes factors
- range of acceptance, rejection and indecision conclusions easily calibrated by simulation
- posterior medians quickly settling near the boundary values of 0 and 1
- potential derivation of a Bayesian *b*-value by looking at the posterior area under the tail of the distribution of the weight

Prior modelling

© Time is ripe for a paradigm shift

- Partly common parameterisation always feasible and hence allows for reference priors
- removal of the absolute prohibition of improper priors in hypothesis testing
- prior on the weight α shows sensitivity that naturally vanishes as the sample size increases
- default value of $a_0 = 0.5$ in the Beta prior

Computing aspects

© Time is ripe for a paradigm shift

- proposal that does not induce additional computational strain
- when algorithmic solutions exist for both models, they can be recycled towards estimating the encompassing mixture
- easier than in standard mixture problems due to common parameters that allow for original MCMC samplers to be turned into proposals
- Gibbs sampling completions useful for assessing potential outliers but not essential to achieve a conclusion about the overall problem