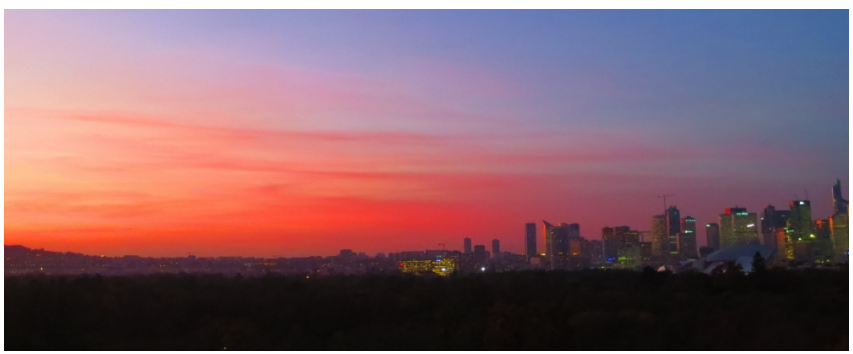
The many nuances of Bayesian testing

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

PHYSTAT seminar, CERN, 19 January 2022



Thanks

- Gilles Celeux
- Nicolas Chopin
- David Frazier
- Florence Forbes
- Kaniav Kamary
- Jean-Michel Marin
- Kerrie Mengersen
- Natesh Pillai
- Judith Rousseau
- Mike Titterington

Outline

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

Testing via mixtures

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 in contrast with confidence sets
- 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)]

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
- and beyond?!

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
- and beyond?!

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 Bayes factor towards supporting a given hypothesis or model, because *not* a Bayesian decision rule with no uncertainty
- similar with posterior probabilities, with (bad) tendency to interpret them as p-values: only report of respective strengths for 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 on Bayes factors falls into the same difficulties as with 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 validating use of *improper priors*: 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 simultaneously ascertain misfit
- no uncertainty assessment attached 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

“Significance tests: one new parameter”

Bayesian testing of hypotheses

Significance tests: one new parameter
- Bayesian tests
- Improper priors for tests

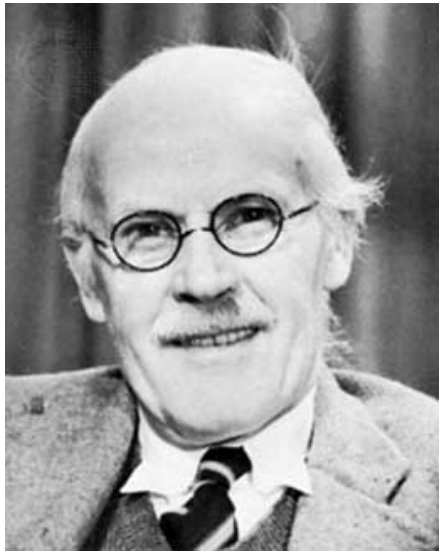
Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking



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 *(Jeffreys,* **ToP***, V, §5.0)*

...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*

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

A (small) 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 or at the very best to make α and β dependent a priori, a choice contradicted by the next paragraph in **ToP**

A (small) 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 or at the very best to make α and β dependent a priori, a choice contradicted by the next paragraph in **ToP**

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

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

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

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

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

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

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).

A major modification

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?!]

> *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 (V, §5.02)*

A major modification

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?!]

> *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 (V, §5.02)*

Point null refurbishment

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 $\pi(\Theta_0) = \rho_0$ and $\pi(\Theta_1) = \rho_1$,

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

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

Point null refurbishment

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 $\pi(\Theta_0) = \rho_0$ and $\pi(\Theta_1) = \rho_1$,

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

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

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

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

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

Connection

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

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

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

Connection

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

A further 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 π_1 or π_2 cannot be coherently normalised **while** the normalisation matters in the Bayes factor

A further 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 π_1 or π_2 cannot be coherently normalised **while** the normalisation matters in the Bayes factor

ToP unaware of the problem?

A. Not entirely, as improper priors keep being used on nuisance parameters

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!

ToP unaware of the problem?

A. Not entirely, as improper priors keep being used on nuisance parameters

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!

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],*

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

and therefore $K = 1$ *(V, §5.2).*

Comments

- **ToP** very imprecise about choice of priors in the setting of tests (despite existence of Bayarri's Jeffreys' conventional partly proper priors)
- **ToP** misses difficulty of improper priors [coherent with earlier stance]
- this problem still generates debates within the B community
- some degree of goodness-of-fit testing but against fixed alternatives

Noninformative proposals

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

what's special about the Bayes factor?!

- Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place?
- 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

[sources: 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 choice of a prior probability of a model
- dire 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

[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

Recall:

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

Recall:

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]

Pseudo-Bayes factors

Idea [LOO]

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* true prior

Pseudo-Bayes factors

Idea [LOO]

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* true prior

Pseudo-Bayes factors

Idea [LOO]

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* true prior

Motivation

- Provides 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 data twice as Aitkin (1991) ["One hardly advances the respect with which statisticians are held in society by making such declarations", D. Lindley]

Motivation

- Provides 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 data twice as Aitkin (1991) ["One hardly advances the respect with which statisticians are held in society by making such declarations", D. Lindley]

Motivation

- Provides 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 data twice as Aitkin (1991) ["One hardly advances the respect with which statisticians are held in society by making such declarations", D. Lindley]

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 $= B^N_{ji} \dfrac{1}{L} \displaystyle\sum_{\ell} B_{ij}(x_{[\ell]})$
  - MIBF $= B^N_{ji}$ med$[B_{ij}(x_{[\ell]})]$
  - GIBF $= B^N_{ji} \exp \dfrac{1}{L} \displaystyle\sum_{\ell} \log B_{ij}(x_{[\ell]})$
- 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

[connection with safe Bayes à la Grünwald (2011)]

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

[connection with safe Bayes à la Grünwald (2011)]

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 posterior and once in p-value)
- needs to be calibrated (back to 0.05?)
- general difficulty in interpreting
- where is the penalty for model complexity?

Jeffreys–Lindley paradox

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox
- Lindley's paradox
- dual versions of the paradox
- Bayesian resolutions

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

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]

Two versions of the paradox

"the weight of Lindley's paradoxical result (...) burdens proponents of the Bayesian practice".

[Lad, 2003]

- official version, opposing frequentist and Bayesian assessments

[Lindley, 1957]

- intra-Bayesian version, blaming vague and improper priors for the Bayes factor misbehaviour:
  if $\pi_1(\cdot|\sigma)$ depends on a scale parameter σ, it is often the case that

$$\mathfrak{B}_{01}(x) \overset{\sigma}{\longrightarrow} +\infty$$

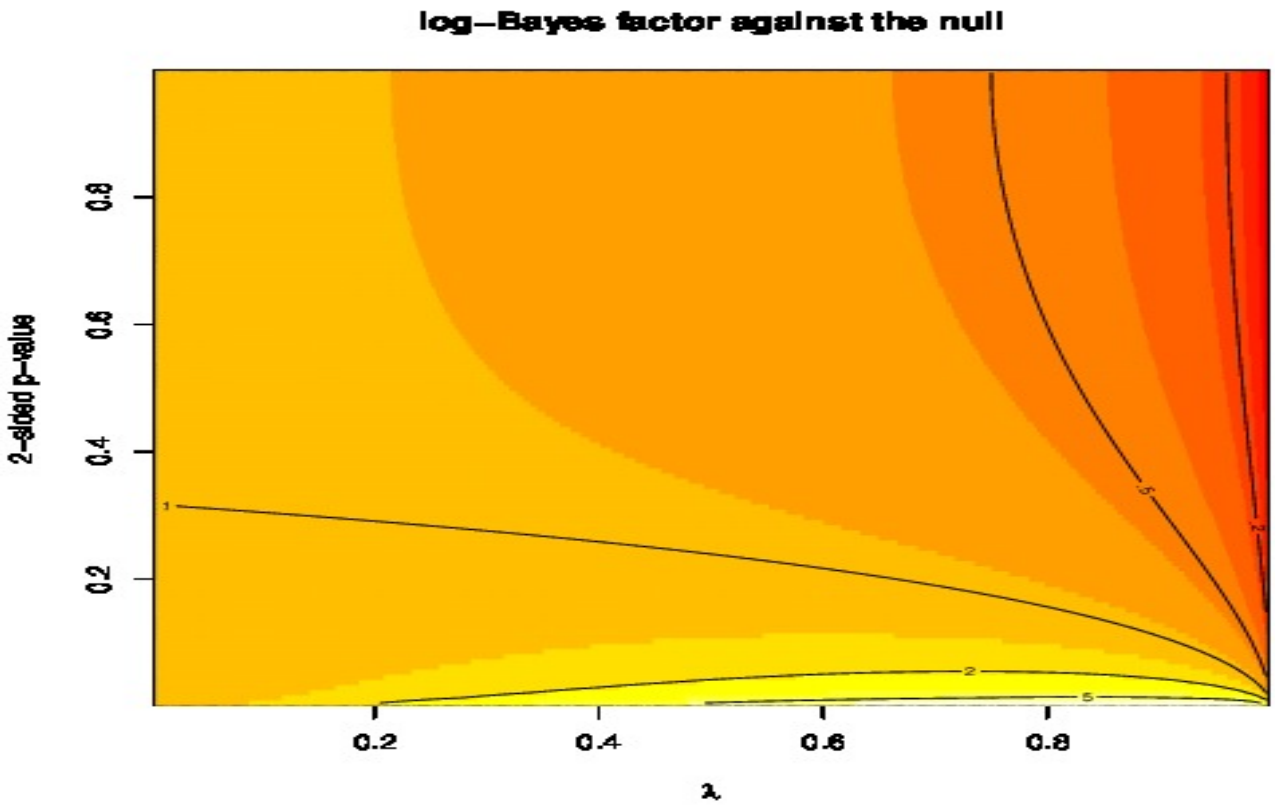
  for a given x, meaning H_0 is always accepted

[Robert, 1992, 2013]

where does it matter?

In the normal case, $Z \sim \mathcal{N}(\theta, 1)$, $\theta \sim \mathcal{N}(0, \alpha^2)$, Bayes factor

$$\mathfrak{B}_{10}(z) = \frac{e^{z^2\alpha^2/(1+\alpha^2)}}{\sqrt{1+\alpha^2}} = \sqrt{1-\lambda}\,\exp\{\lambda z^2/2\}$$



Evacuation of the first version

Two paradigms [(𝔟) versus (𝔣)]

- one (𝔟) operates on the parameter space Θ, while the other (𝔣) is produced from the sample space
- one (𝔣) relies solely on the point-null hypothesis H_0 and the corresponding sampling distribution, while the other (𝔟) opposes H_0 to a (predictive) marginal version of H_1
- one (𝔣) could reject *"a hypothesis that may be true (...) because it has not predicted observable results that have not occurred" (Jeffreys,* **ToP***, VII, §7.2)* while the other (𝔟) conditions upon the observed value x_{obs}
- one (𝔣) cannot agree with the likelihood principle, while the other (𝔟) is almost uniformly in agreement with it
- one (𝔣) resorts to an arbitrary fixed bound α on the p-value, while the other (𝔟) refers to the (default) boundary probability of $1/2$

Nothing's wrong with the second version

- n, prior's scale factor: prior variance n times larger than the observation variance and *when n goes to ∞, Bayes factor goes to ∞ no matter what the observation is*
- n becomes what Lindley (1957) calls *"a measure of lack of conviction about the null hypothesis"*
- when prior diffuseness under H_1 increases, only relevant information becomes that θ could be equal to θ_0, and this overwhelms any evidence to the contrary contained in the data
- mass of the prior distribution in the vicinity of any fixed neighbourhood of the null hypothesis vanishes to zero under H_1

© **deep coherence in the outcome: being indecisive about the alternative hypothesis means we should not chose it**

Nothing's wrong with the second version

- n, prior's scale factor: prior variance n times larger than the observation variance and *when n goes to ∞, Bayes factor goes to ∞ no matter what the observation is*
- n becomes what Lindley (1957) calls *"a measure of lack of conviction about the null hypothesis"*
- when prior diffuseness under H_1 increases, only relevant information becomes that θ could be equal to θ_0, and this overwhelms any evidence to the contrary contained in the data
- mass of the prior distribution in the vicinity of any fixed neighbourhood of the null hypothesis vanishes to zero under H_1

© **deep coherence in the outcome: being indecisive about the alternative hypothesis means we should not chose it**

On some resolutions of the second version

- 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,
- use of 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 second version

- 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]

- use of 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 second version

- 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]

- use of 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 second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of 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 second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of 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 second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of 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 second version

- use of pseudo-Bayes factors, fractional Bayes factors, &tc,
- use of *identical* improper priors on nuisance parameters,
- use of 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]

Deviance (information criterion)

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

Testing via mixtures

DIC as in Dayesian?

Deviance defined by

$$D(\theta) = -2\log(p(\mathbf{y}|\theta))\,,$$

Effective number of parameters computed as

$$p_D = \bar{D} - D(\bar{\theta})\,,$$

with $\bar{D}$ posterior expectation of D and $\bar{\theta}$ estimate of θ
Deviance information criterion (DIC) defined by

$$\begin{aligned}\mathrm{DIC} &= p_D + \bar{D} \\ &= D(\bar{\theta}) + 2p_D\end{aligned}$$

Models with smaller DIC better supported by the data

[Spiegelhalter et al., 2002]

"thou shalt not use the data twice"

The data is used twice in the DIC method:

1. y used **once** to produce the posterior $\pi(\theta|y)$, and the associated estimate, $\tilde{\theta}(y)$
2. y used **a second time** to compute the posterior expectation of the *observed* likelihood $p(y|\theta)$,

$$\int \log p(y|\theta)\pi(\mathrm{d}\theta|y) \propto \int \log p(y|\theta)p(y|\theta)\pi(\mathrm{d}\theta)\,,$$

DIC for missing data models

Framework of missing data models

$$f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \boldsymbol{z}|\theta) d\boldsymbol{z}\,,$$

with observed data $\mathbf{y} = (y_1, \ldots, y_n)$ and corresponding *missing data* by $\boldsymbol{z} = (z_1, \ldots, z_n)$

How do we define DIC in such settings?

DIC for missing data models

Framework of missing data models

$$f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \boldsymbol{z}|\theta) d\boldsymbol{z}\,,$$

with observed data $\mathbf{y} = (y_1, \ldots, y_n)$ and corresponding *missing data* by $\boldsymbol{z} = (z_1, \ldots, z_n)$

How do we define DIC in such settings?

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_1 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\mathbb{E}_\theta\left[\theta|\mathbf{y}\right])$$

   often a poor choice in case of unidentifiability
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_2 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}))\,.$$

   which uses posterior mode instead
2. complete DICs based on $f(\mathbf{y},\mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z},\theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs

$$\text{DIC}_3 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log \widehat{f}(\mathbf{y})\,,$$

   which instead relies on the MCMC density estimate
2. complete DICs based on $f(\mathbf{y},\mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z},\theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

$$\begin{aligned} \text{DIC}_4 &= \mathbb{E}_{\mathbf{Z}}\left[\text{DIC}(\mathbf{y}, \mathbf{Z})|\mathbf{y}\right] \\ &= -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\mathbb{E}_{\theta}[\theta|\mathbf{y}, \mathbf{Z}])|\mathbf{y}\right] \end{aligned}$$

3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

$$\text{DIC}_5 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}, \widehat{\mathbf{z}}(\mathbf{y})|\widehat{\theta}(\mathbf{y}))\,,$$

   using $\mathbf{Z}$ as an additional parameter
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$

   $$\text{DIC}_6 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\widehat{\theta}(\mathbf{y}))|\mathbf{y}, \widehat{\theta}(\mathbf{y})\right].$$

   in analogy with EM, $\widehat{\theta}$ being an EM fixed point
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

$$\text{DIC}_7 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\mathbf{z}}(\mathbf{y}), \widehat{\theta}(\mathbf{y}))\,,$$

using MAP estimates

[Celeux et al., BA, 2006]

how many DICs can you fit in a mixture?

Q: How many giraffes can you fit in a VW bug?
A: None, the elephants are in there.

1. observed DICs
2. complete DICs based on $f(\mathbf{y}, \mathbf{z}|\theta)$
3. conditional DICs based on $f(\mathbf{y}|\mathbf{z}, \theta)$

$$\text{DIC}_8 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z}, \widehat{\theta}(\mathbf{y}, \mathbf{Z}))|\mathbf{y}\right] ,$$

conditioning first on $\mathbf{Z}$ and then integrating over $\mathbf{Z}$ conditional on $\mathbf{y}$

[Celeux et al., BA, 2006]

Galactic DICs

Example of the galaxy mixture dataset

K	DIC_2 (p_{D2})	DIC_3 (p_{D3})	DIC_4 (p_{D4})	DIC_5 (p_{D5})	DIC_6 (p_{D6})	DIC_7 (p_{D7})	DIC_8 (p_{D8})
2	453	451	502	705	501	417	410
	(5.56)	(3.66)	(5.50)	(207.88)	(4.48)	(11.07)	(4.09)
3	440	436	461	622	471	378	372
	(9.23)	(4.94)	(6.40)	(167.28)	(15.80)	(13.59)	(7.43)
4	446	439	473	649	482	388	382
	(11.58)	(5.41)	(7.52)	(183.48)	(16.51)	(17.47)	(11.37)
5	447	442	485	658	511	395	390
	(10.80)	(5.48)	(7.58)	(180.73)	(33.29)	(20.00)	(15.15)
6	449	444	494	676	532	407	398
	(11.26)	(5.49)	(8.49)	(191.10)	(46.83)	(28.23)	(19.34)
7	460	446	508	700	571	425	409
	(19.26)	(5.83)	(8.93)	(200.35)	(71.26)	(40.51)	(24.57)

questions

- what is the behaviour of DIC under model mispecification?
- is there an absolute scale to the DIC values, i.e. when is a difference in DICs significant?
- how can DIC handle small n's versus p's?
- should p_D be defined as $\text{var}(D|\mathbf{y})/2$ [Gelman's suggestion]?
- is WAIC (Gelman and Vehtari, 2013; Watanabe, 2017) making a difference for being based on expected posterior predictive?

In an era of complex models, is DIC applicable?

[Robert, 2013]

questions

- what is the behaviour of DIC under model mispecification?
- is there an absolute scale to the DIC values, i.e. when is a difference in DICs significant?
- how can DIC handle small n's versus p's?
- should p_D be defined as $\text{var}(D|\mathbf{y})/2$ [Gelman's suggestion]?
- is WAIC (Gelman and Vehtari, 2013; Watanabe, 2017) making a difference for being based on expected posterior predictive?

In an era of complex models, is DIC applicable?

[Robert, 2013]

Testing under incomplete information

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

Testing via mixtures

Likelihood-free settings

Cases when the likelihood function $f(\mathbf{y}|\theta)$ is unavailable (in analytic and numerical senses) and when the completion step

$$f(\mathbf{y}|\theta) = \int_{\mathcal{Z}} f(\mathbf{y}, \boldsymbol{z}|\theta)\, \mathrm{d}\boldsymbol{z}$$

is impossible or too costly because of the dimension of $\boldsymbol{z}$

© MCMC cannot be implemented!

The ABC method

Bayesian setting: target is $\pi(\theta)f(x|\theta)$

When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, \mathbf{z} \sim f(\mathbf{z}|\theta')\,,$$

until the auxiliary variable $\mathbf{z}$ is equal to the observed value, $\mathbf{z} = \mathbf{y}$.

[Tavaré et al., 1997]

The ABC method

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, \boldsymbol{z} \sim f(\boldsymbol{z}|\theta')\,,$$

until the auxiliary variable $\boldsymbol{z}$ is equal to the observed value, $\boldsymbol{z} = \mathbf{y}$.

[Tavaré et al., 1997]

The ABC method

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, \boldsymbol{z} \sim f(\boldsymbol{z}|\theta')\,,$$

until the auxiliary variable $\boldsymbol{z}$ is equal to the observed value, $\boldsymbol{z} = \mathbf{y}$.

[Tavaré et al., 1997]

A as A...pproximative

When y is a continuous random variable, strict equality $\boldsymbol{z} = \boldsymbol{y}$ is replaced with a tolerance zone

$$\rho(\boldsymbol{y}, \boldsymbol{z}) \leqslant \epsilon$$

where ρ is a distance

Output distributed from

$$\pi(\theta)\, P_\theta\{\rho(\boldsymbol{y}, \boldsymbol{z}) < \epsilon\} \overset{\text{def}}{\propto} \pi(\theta | \rho(\boldsymbol{y}, \boldsymbol{z}) < \epsilon)$$

[Pritchard et al., 1999]

A as A...pproximative

When y is a continuous random variable, strict equality $\boldsymbol{z} = \boldsymbol{y}$ is replaced with a tolerance zone

$$\rho(\boldsymbol{y}, \boldsymbol{z}) \leqslant \epsilon$$

where ρ is a distance
Output distributed from

$$\pi(\theta)\, P_\theta\{\rho(\boldsymbol{y}, \boldsymbol{z}) < \epsilon\} \overset{\text{def}}{\propto} \pi(\theta | \rho(\boldsymbol{y}, \boldsymbol{z}) < \epsilon)$$

[Pritchard et al., 1999]

ABC algorithm

In most implementations, further degree of A...pproximation:

Algorithm 1 Likelihood-free rejection sampler

for $i = 1$ to N **do**
  repeat
    generate θ' from the prior distribution $\pi(\cdot)$
    generate $\boldsymbol{z}$ from the likelihood $f(\cdot|\theta')$
  until $\rho\{\eta(\boldsymbol{z}), \eta(\mathbf{y})\} \leqslant \epsilon$
  set $\theta_i = \theta'$
end for

where $\eta(\mathbf{y})$ defines a (not necessarily sufficient) statistic

Which summary $\eta(\cdot)$?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., LDA)
- may gather several non-B point estimates
- can learn about efficient combination

[Estoup et al., 2012, Genetics]

Which summary $\eta(\cdot)$?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., LDA)
- may gather several non-B point estimates
- can learn about efficient combination

[Estoup et al., 2012, Genetics]

Which summary $\eta(\cdot)$?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation
- may be imposed for external/practical reasons (e.g., LDA)
- may gather several non-B point estimates
- can learn about efficient combination

[Estoup et al., 2012, Genetics]

Generic ABC for model choice

Algorithm 2 Likelihood-free model choice sampler (ABC-MC)

for $t = 1$ to T **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate θ_m from the prior $\pi_m(\theta_m)$
    Generate z from the model $f_m(z|\theta_m)$
  until $\rho\{\eta(z), \eta(y)\} < \epsilon$
  Set $m^{(t)} = m$ and $\theta^{(t)} = \theta_m$
end for

[Grelaud et al., 2009]

Formalised framework

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $\eta(y)$ consistent?

Note: © drawn on $\eta(y)$ through $\mathfrak{B}^{\eta}_{12}(y)$ necessarily differs from © drawn on y through $\mathfrak{B}_{12}(y)$

Formalised framework

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $\eta(y)$ consistent?

Note: © drawn on $\eta(\mathbf{y})$ through $\mathfrak{B}^{\eta}_{12}(\mathbf{y})$ necessarily differs from © drawn on $\mathbf{y}$ through $\mathfrak{B}_{12}(\mathbf{y})$

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

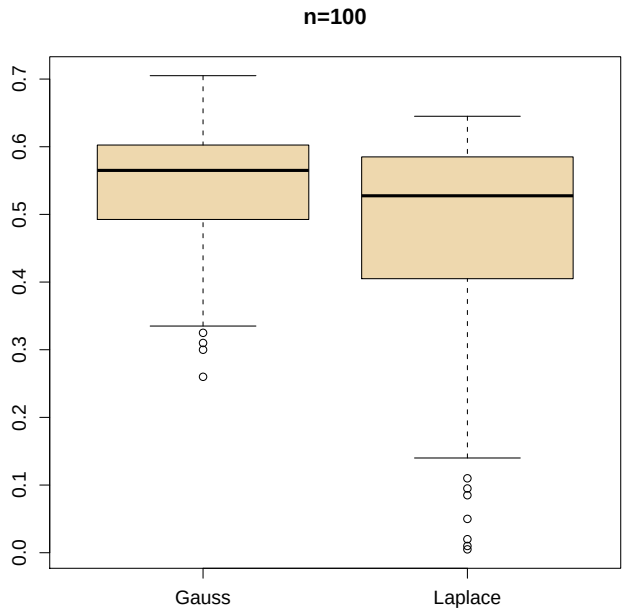
Four possible statistics

1. sample mean $\overline{\mathbf{y}}$ (sufficient for $\mathfrak{M}_1$ if not $\mathfrak{M}_2$);
2. sample median $\text{med}(\mathbf{y})$ (insufficient);
3. sample variance $\text{var}(\mathbf{y})$ (ancillary);
4. median absolute deviation $\text{mad}(\mathbf{y}) = \text{med}(\mathbf{y} - \text{med}(\mathbf{y}))$;

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

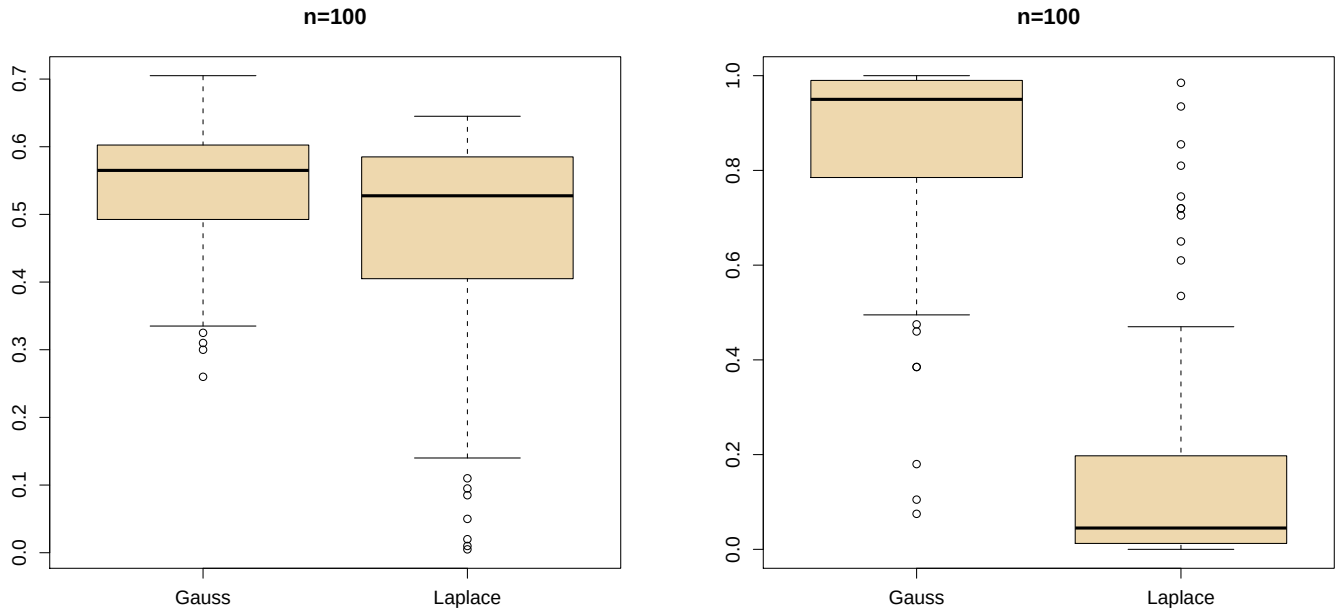
Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).



A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).



Consistency theorem

If P^n belongs to one of the two models and if $\mu_0 = \mathbb{E}[\boldsymbol{\eta}]$ cannot be attained by the other one :

$$\begin{aligned} 0 &= \min\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right) \\ &< \max\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right), \end{aligned}$$

then the Bayes factor $\mathfrak{B}_{12}^{\boldsymbol{\eta}}$ is consistent

Changing the testing perspective

Bayesian testing of hypotheses

Significance tests: one new parameter

Noninformative solutions

Jeffreys-Lindley paradox

Deviance (information criterion)

Testing under incomplete information

Posterior predictive checking

Testing via mixtures

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

Poisson/Geometric

- choice betwen Poisson $\mathcal{P}(\lambda)$ and Geometric $\mathcal{G}eo(p)$ distribution
- mixture with common parameter λ

$$\mathfrak{M}_\alpha : \ \alpha\mathcal{P}(\lambda) + (1-\alpha)\mathcal{G}eo(1/1+\lambda)$$

Allows for Jeffreys prior since resulting posterior is proper

- independent Metropolis–within–Gibbs with proposal distribution on λ equal to Poisson posterior (with acceptance rate larger than 75%)

Poisson/Geometric

- choice betwen Poisson $\mathcal{P}(\lambda)$ and Geometric $\mathcal{G}eo(p)$ distribution
- mixture with common parameter λ

$$\mathfrak{M}_\alpha : \ \alpha\mathcal{P}(\lambda) + (1-\alpha)\mathcal{G}eo(1/1+\lambda)$$

Allows for Jeffreys prior since resulting posterior is proper

- independent Metropolis–within–Gibbs with proposal distribution on λ equal to Poisson posterior (with acceptance rate larger than 75%)

Poisson/Geometric

- choice betwen Poisson $\mathcal{P}(\lambda)$ and Geometric $\mathcal{G}eo(p)$ distribution
- mixture with common parameter λ

$$\mathfrak{M}_\alpha : \ \alpha\mathcal{P}(\lambda) + (1-\alpha)\mathcal{G}eo(1/1+\lambda)$$

Allows for Jeffreys prior since resulting posterior is proper

- independent Metropolis–within–Gibbs with proposal distribution on λ equal to Poisson posterior (with acceptance rate larger than 75%)

Beta prior

When $\alpha \sim \mathcal{B}e(a_0, a_0)$ prior, full conditional posterior

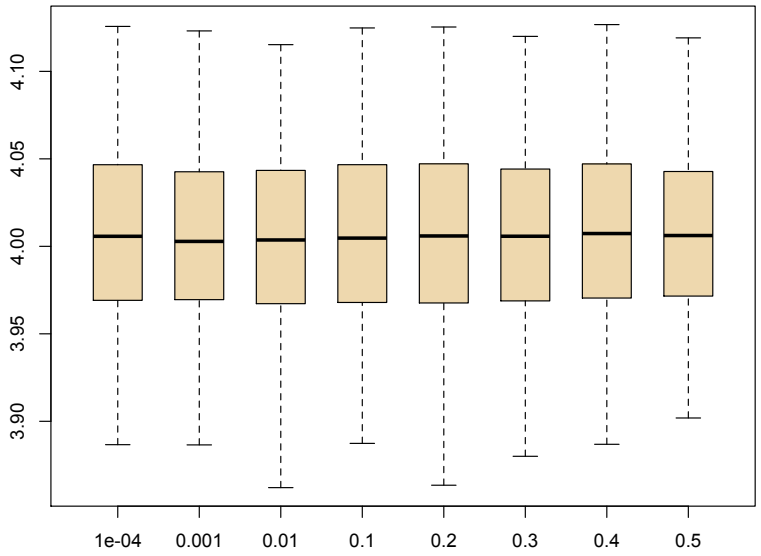
$$\alpha \sim \mathcal{B}e(n_1(\zeta) + a_0, n_2(\zeta) + a_0)$$

Exact Bayes factor opposing Poisson and Geometric

$$\mathfrak{B}_{12} = n^{n\bar{x}_n} \prod_{i=1}^{n} x_i! \, \Gamma\left(n + 2 + \sum_{i=1}^{n} x_i\right) / \Gamma(n+2)$$

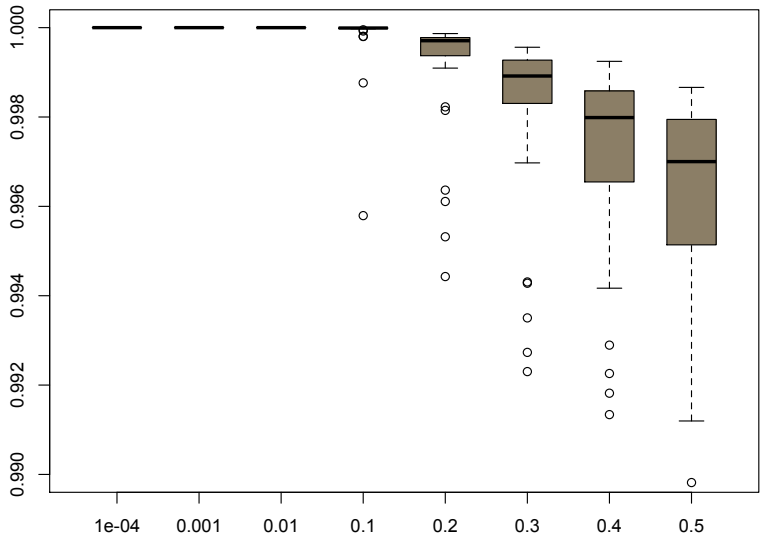
although arbitrary from a purely mathematical viewpoint

Parameter estimation



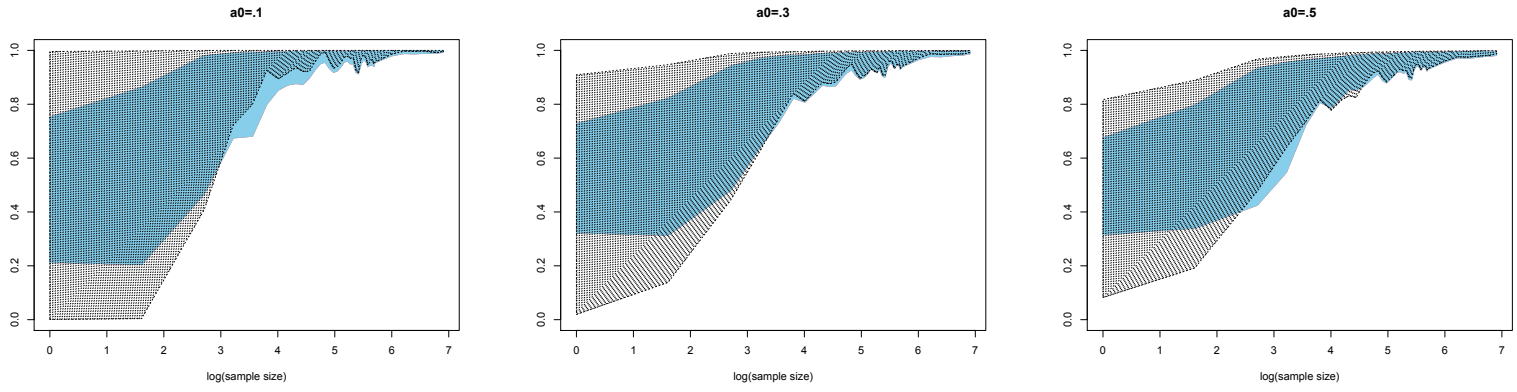
Posterior means of λ and medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Parameter estimation



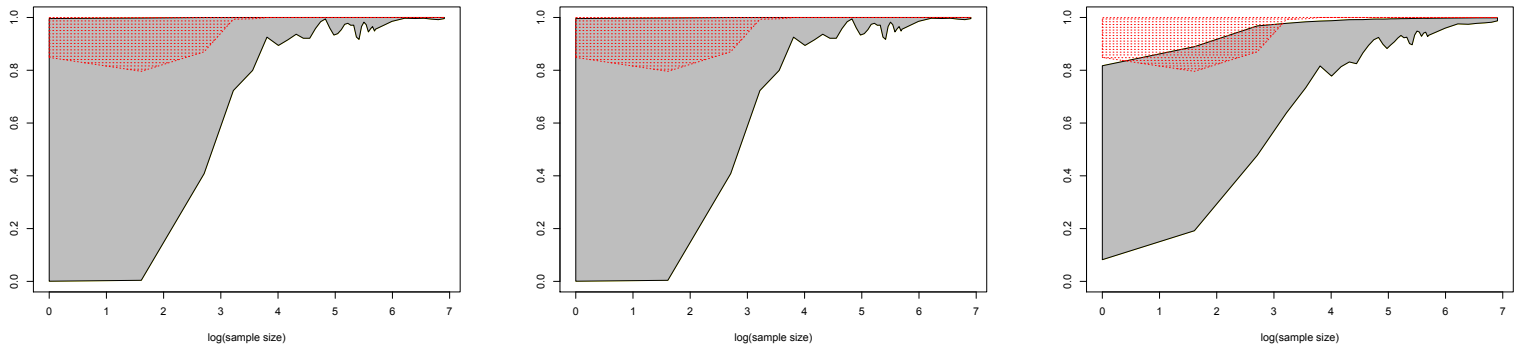
Posterior means of λ and medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Consistency



Posterior means *(sky-blue)* and medians *(grey-dotted)* of α, over 100 Poisson $\mathcal{P}(4)$ datasets for sample sizes from 1 to 1000.

Behaviour of Bayes factor



Comparison between $\mathbb{P}(\mathfrak{M}_1|x)$ *(red dotted area)* and posterior medians of α *(grey zone)* for 100 Poisson $\mathcal{P}(4)$ datasets with sample sizes n between 1 and 1000, for $a_0 = .001, .1, .5$

Normal-normal comparison

- comparison of a normal $\mathcal{N}(\theta_1, 1)$ with a normal $\mathcal{N}(\theta_2, 2)$ distribution
- mixture with identical location parameter θ $\alpha\mathcal{N}(\theta, 1) + (1-\alpha)\mathcal{N}(\theta, 2)$
- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper
- Reference (improper) Bayes factor

$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

Normal-normal comparison

- comparison of a normal $\mathcal{N}(\theta_1, 1)$ with a normal $\mathcal{N}(\theta_2, 2)$ distribution
- mixture with identical location parameter θ $\alpha\mathcal{N}(\theta, 1) + (1-\alpha)\mathcal{N}(\theta, 2)$
- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper
- Reference (improper) Bayes factor

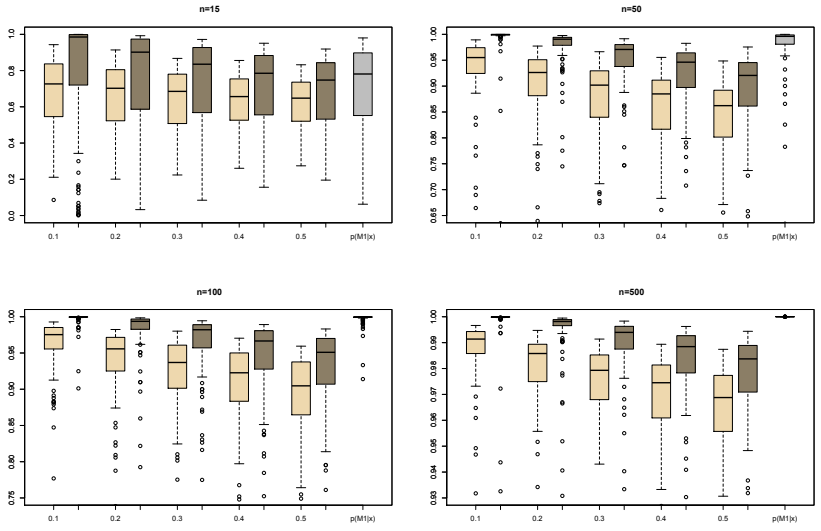
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

Normal-normal comparison

- comparison of a normal $\mathcal{N}(\theta_1, 1)$ with a normal $\mathcal{N}(\theta_2, 2)$ distribution
- mixture with identical location parameter θ
  $\alpha\mathcal{N}(\theta, 1) + (1-\alpha)\mathcal{N}(\theta, 2)$
- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper
- Reference (improper) Bayes factor

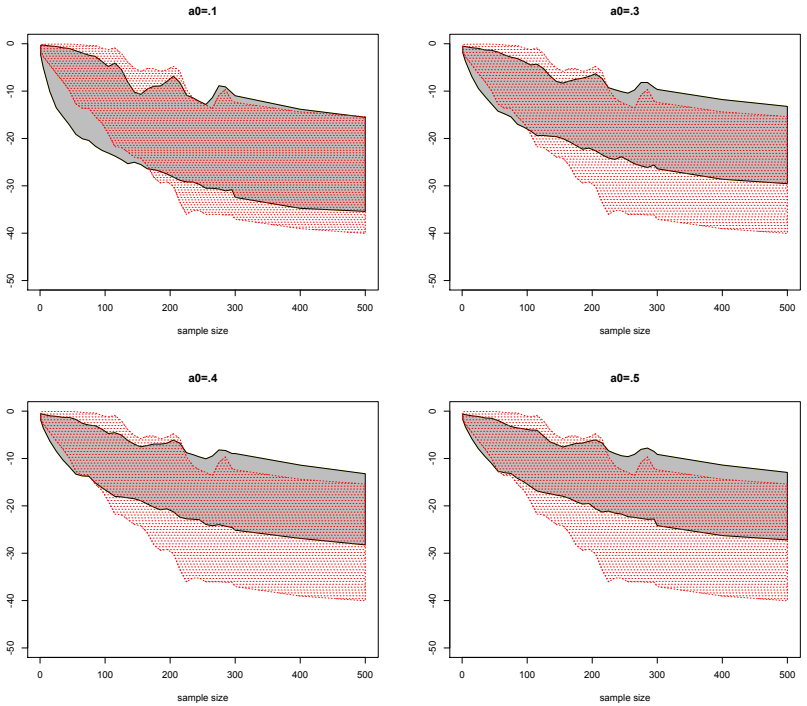
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

Consistency



Posterior means *(wheat)* and medians of α *(dark wheat)*, compared with posterior probabilities of $\mathfrak{M}_0$ *(gray)* for a $\mathcal{N}(0,1)$ sample, derived from 100 datasets for sample sizes equal to $15, 50, 100, 500$. Each posterior approximation is based on 10^4 MCMC iterations.

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.

Comments

- convergence to one boundary value as sample size n grows
- impact of hyperarameter a_0 slowly vanishes as n increases, but present for moderate sample sizes
- when simulated sample is neither from $\mathcal{N}(\theta_1, 1)$ nor from $\mathcal{N}(\theta_2, 2)$, behaviour of posterior varies, depending on which distribution is closest

Logit or Probit?

- binary dataset, R dataset about diabetes in 200 Pima Indian women with body mass index as explanatory variable
- comparison of logit and probit fits could be suitable. We are thus comparing both fits via our method

$$\mathfrak{M}_1 : y_i \mid \mathbf{x}^i, \theta_1 \sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i \theta_1)}{1 + \exp(\mathbf{x}^i \theta_1)}$$

$$\mathfrak{M}_2 : y_i \mid \mathbf{x}^i, \theta_2 \sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i \theta_2)$$

Common parameterisation

Local reparameterisation strategy that rescales parameters of the probit model $\mathfrak{M}_2$ so that the MLE's of both models coincide.

[Choudhuty et al., 2007]

$$\Phi(\mathbf{x}^i\theta_2) \approx \frac{\exp(k\mathbf{x}^i\theta_2)}{1+\exp(k\mathbf{x}^i\theta_2)}$$

and use best estimate of k to bring both parameters into coherency

$$(k_0, k_1) = (\widehat{\theta_{01}}/\widehat{\theta_{02}}, \widehat{\theta_{11}}/\widehat{\theta_{12}})\,,$$

reparameterise $\mathfrak{M}_1$ and $\mathfrak{M}_2$ as

$$\begin{aligned}
\mathfrak{M}_1 &: y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta)}{1+\exp(\mathbf{x}^i\theta)} \\
\mathfrak{M}_2 &: y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i(\kappa^{-1}\theta))\,,
\end{aligned}$$

with $\kappa^{-1}\theta = (\theta_0/k_0, \theta_1/k_1)$.

Prior modelling

Under default g-prior

$$\theta \sim \mathcal{N}_2(0, n(X^{\mathsf{T}}X)^{-1})$$

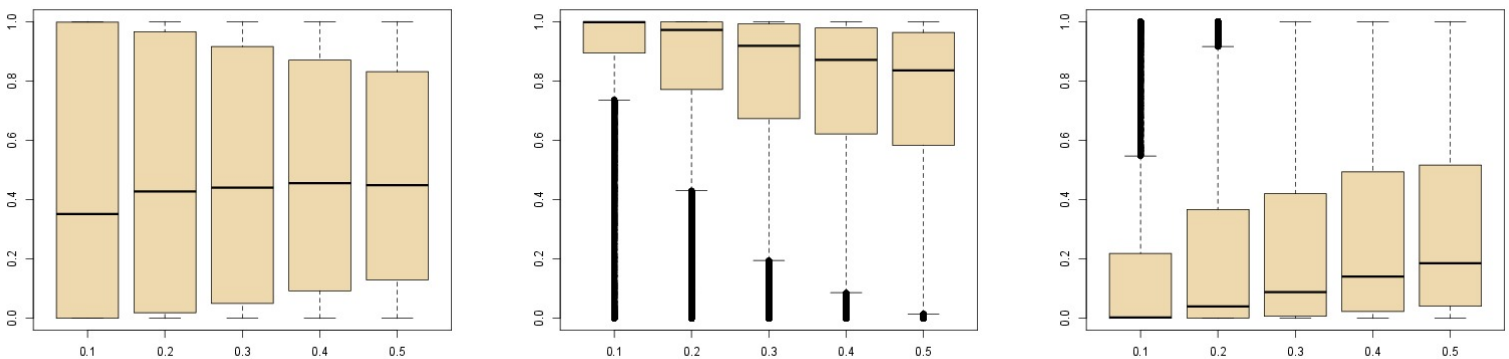
full conditional posterior distributions given allocations

$$\pi(\theta \mid \mathbf{y}, X, \zeta) \propto \frac{\exp\left\{\sum_i \mathbb{I}_{\zeta_i=1} y_i \mathbf{x}^i\theta\right\}}{\prod_{i;\zeta_i=1}[1+\exp(\mathbf{x}^i\theta)]} \exp\left\{-\theta^T(X^TX)\theta/2n\right\}$$
$$\times \prod_{i;\zeta_i=2} \Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{(1-y_i)}$$

hence posterior distribution clearly defined

Results

		Logistic		**Probit**	
a_0	α	θ_0	θ_1	$\frac{\theta_0}{k_0}$	$\frac{\theta_1}{k_1}$
.1	.352	-4.06	.103	-2.51	.064
.2	.427	-4.03	.103	-2.49	.064
.3	.440	-4.02	.102	-2.49	.063
.4	.456	-4.01	.102	-2.48	.063
.5	.449	-4.05	.103	-2.51	.064



Histograms of posteriors of α in favour of logistic model where $a_0 = .1, .2, .3, .4, .5$ for (a) Pima dataset, (b) Data from logistic model, (c) Data from probit model

Survival analysis

Testing hypothesis that data comes from a

1. log-Normal(ϕ, κ^2),
2. Weibull(α, λ), or
3. log-Logistic(γ, δ)

distribution

Corresponding mixture given by the density

$$\alpha_1 \exp\{-(\log x - \phi)^2/2\kappa^2\}/\sqrt{2\pi}x\kappa +$$
$$\alpha_2 \frac{\alpha}{\lambda} \exp\{-(x/\lambda)^\alpha\}((x/\lambda)^{\alpha-1} +$$
$$\alpha_3(\delta/\gamma)(x/\gamma)^{\delta-1}/(1+(x/\gamma)^\delta)^2$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$

Survival analysis

Testing hypothesis that data comes from a

1. log-Normal(ϕ, κ^2),
2. Weibull(α, λ), or
3. log-Logistic(γ, δ)

distribution

Corresponding mixture given by the density

$$\alpha_1 \exp\{-(\log x - \phi)^2/2\kappa^2\}/\sqrt{2\pi}x\kappa + \alpha_2 \frac{\alpha}{\lambda} \exp\{-(x/\lambda)^\alpha\}((x/\lambda)^{\alpha-1} + \alpha_3(\delta/\gamma)(x/\gamma)^{\delta-1}/(1+(x/\gamma)^\delta)^2$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$

Reparameterisation

Looking for common parameter(s):

$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

where $\gamma \approx 0.5772$ is Euler-Mascheroni constant.

Allows for a noninformative prior on the common location scale parameter,

$$\pi(\phi, \sigma^2) = 1/\sigma^2$$

Reparameterisation

Looking for common parameter(s):

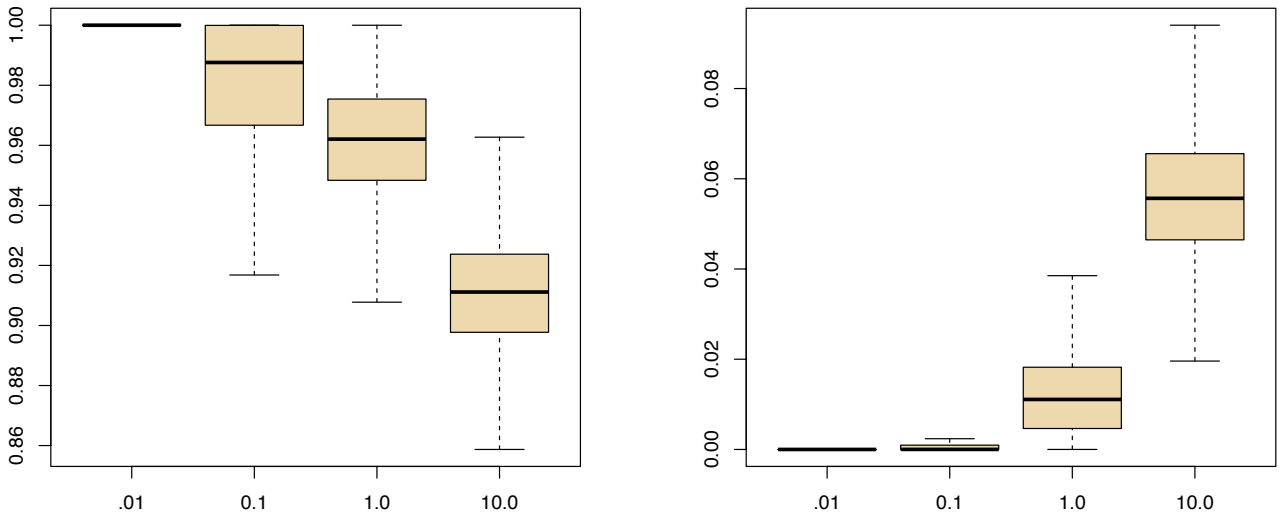
$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

where $\gamma \approx 0.5772$ is Euler-Mascheroni constant.

Allows for a noninformative prior on the common location scale parameter,

$$\pi(\phi, \sigma^2) = 1/\sigma^2$$

Recovery



Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii: truth = Normal *(left panel)*, truth = Gumbel *(right panel)*, a_0=0.01, 0.1, 1.0, 10.0 *(from left to right in each panel)* and $n = 10,000$ simulated observations.

Asymptotic consistency

Posterior consistency holds for mixture testing procedure [under minor conditions]

Two different cases

- the two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, are well separated
- model $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$.

Asymptotic consistency

Posterior consistency holds for mixture testing procedure [under minor conditions]

Two different cases

- the two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, are well separated
- model $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$.

Posterior concentration rate

Let π be the prior and $\mathbf{x}^n = (x_1, \cdots, x_n)$ a sample with true density f^*

proposition

Assume that, for all $c > 0$, there exist $\Theta_n \subset \Theta_1 \times \Theta_2$ and $B > 0$ such that

$$\pi\left[\Theta_n^c\right] \leqslant n^{-c}, \quad \Theta_n \subset \{\|\theta_1\| + \|\theta_2\| \leqslant n^B\}$$

and that there exist $H \geqslant 0$ and $L, \delta > 0$ such that, for $j = 1, 2$,

$$\sup_{\theta, \theta' \in \Theta_n} \|f_{j,\theta_j} - f_{j,\theta_j'}\|_1 \leqslant L n^H \|\theta_j - \theta_j'\|, \quad \theta = (\theta_1, \theta_2),\ \theta' = (\theta_1', \theta_2'),$$

$$\forall \|\theta_j - \theta_j^*\| \leqslant \delta; \quad KL(f_{j,\theta_j}, f_{j,\theta_j^*}) \lesssim \|\theta_j - \theta_j^*\|.$$

Then, when $f^* = f_{\theta^*,\alpha^*}$, with $\alpha^* \in [0, 1]$, there exists $M > 0$ such that

$$\pi\left[(\alpha, \theta); \|f_{\theta,\alpha} - f^*\|_1 > M\sqrt{\log n / n} | \mathbf{x}^n\right] = o_p(1).$$

Separated models

Assumption: Models are separated, i.e. identifiability holds:

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta'$$

Further

$$\inf_{\theta_1 \in \Theta_1} \inf_{\theta_2 \in \Theta_2} \| f_{1,\theta_1} - f_{2,\theta_2} \|_1 > 0$$

and, for $\theta_j^* \in \Theta_j$, if P_{θ_j} weakly converges to $P_{\theta_j^*}$, then

$$\theta_j \longrightarrow \theta_j^*$$

in the Euclidean topology

Separated models

Assumption: Models are separated, i.e. identifiability holds:

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

theorem

Under above assumptions, then for all $\epsilon > 0$,

$$\pi\left[|\alpha - \alpha^*| > \epsilon | \mathbf{x}^n\right] = o_p(1)$$

Separated models

Assumption: Models are separated, i.e. identifiability holds:

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta'$$

theorem

If

- $\theta_j \longrightarrow f_{j,\theta_j}$ is $\mathcal{C}^2$ around θ_j^*, $j = 1,2$,
- $f_{1,\theta_1^*} - f_{2,\theta_2^*}, \nabla f_{1,\theta_1^*}, \nabla f_{2,\theta_2^*}$ are linearly independent in y and
- there exists $\delta > 0$ such that

$$\nabla f_{1,\theta_1^*}, \nabla f_{2,\theta_2^*}, \sup_{|\theta_1 - \theta_1^*| < \delta} |D^2 f_{1,\theta_1}|, \sup_{|\theta_2 - \theta_2^*| < \delta} |D^2 f_{2,\theta_2}| \in L_1$$

then

$$\pi\left[|\alpha - \alpha^*| > M\sqrt{\log n / n} \middle| x^n\right] = o_p(1).$$

Separated models

Assumption: Models are separated, i.e. identifiability holds:

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j^{'}, j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha^{'}, \quad \theta = \theta^{'}$$

theorem allows for interpretation of α under the posterior: If data $\boldsymbol{x}^n$ is generated from model $\mathfrak{M}_1$ then posterior on α concentrates around $\alpha = 1$

Embedded case

Here $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$, i.e.

$$\theta_2 = (\theta_1, \psi) \quad \text{and} \quad \theta_2 = (\theta_1, \psi_0 = 0)$$

corresponds to $f_{2,\theta_2} \in \mathfrak{M}_1$
Same posterior concentration rate

$$\sqrt{\log n / n}$$

for estimating α when $\alpha^* \in (0, 1)$ and $\psi^* \neq 0$.

Null case

- Case where $\psi^* = 0$, i.e., f^* is in model $\mathfrak{M}_1$
- Two possible paths to approximate f^*: either α goes to 1 (path 1) or ψ goes to 0 (path 2)
- New identifiability condition: $P_{\theta,\alpha} = P^*$ only if

$$\alpha = 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leqslant 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, 0)$$

Prior

$$\pi(\alpha, \theta) = \pi_\alpha(\alpha)\pi_1(\theta_1)\pi_\psi(\psi), \quad \theta_2 = (\theta_1, \psi)$$

with common (prior on) θ_1

Null case

- Case where $\psi^* = 0$, i.e., f^* is in model $\mathfrak{M}_1$
- Two possible paths to approximate f^*: either α goes to 1 (path 1) or ψ goes to 0 (path 2)
- New identifiability condition: $P_{\theta,\alpha} = P^*$ only if

$$\alpha = 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leqslant 1,\ \theta_1 = \theta_1^*,\ \theta_2 = (\theta_1^*, 0)$$

Prior

$$\pi(\alpha, \theta) = \pi_\alpha(\alpha)\pi_1(\theta_1)\pi_\psi(\psi), \quad \theta_2 = (\theta_1, \psi)$$

with common (prior on) θ_1

Assumptions

[B1] *Regularity*: Assume that $\theta_1 \longrightarrow f_{1,\theta_1}$ and $\theta_2 \longrightarrow f_{2,\theta_2}$ are 3 times continuously differentiable and that

$$F^*\left(\frac{\bar{f}^3_{1,\theta_1^*}}{\underline{f}^3_{1,\theta_1^*}}\right) < +\infty, \quad \bar{f}_{1,\theta_1^*} = \sup_{|\theta_1-\theta_1^*|<\delta} f_{1,\theta_1}, \quad \underline{f}_{1,\theta_1^*} = \inf_{|\theta_1-\theta_1^*|<\delta} f_{1,\theta_1}$$

$$F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|\nabla f_{1,\theta_1^*}|^3}{\underline{f}^3_{1,\theta_1^*}}\right) < +\infty, \quad F^*\left(\frac{|\nabla f_{1,\theta_1^*}|^4}{f^4_{1,\theta_1^*}}\right) < +\infty,$$

$$F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|D^2 f_{1,\theta_1^*}|^2}{\underline{f}^2_{1,\theta_1^*}}\right) < +\infty, \quad F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|D^3 f_{1,\theta_1^*}|}{\underline{f}_{1,\theta_1^*}}\right) < +\infty$$

Assumptions

[B2] *Integrability*: There exists

$$\mathcal{S}_0 \subset \mathcal{S} \cap \{|\psi| > \delta_0\}$$

for some positive δ_0 and satisfying $\text{Leb}(\mathcal{S}_0) > 0$, and such that for all $\psi \in \mathcal{S}_0$,

$$F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f_{2,\theta_1,\psi}}{f^4_{1,\theta_1^*}} \right) < +\infty, \quad F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f^3_{2,\theta_1,\psi}}{\underline{f}^3_{1,\theta 1^*}} \right) < +\infty,$$

Assumptions

[B3] *Stronger identifiability*: Set

$$\nabla f_{2,\theta_1^*,\psi^*}(x) = \left(\nabla_{\theta_1} f_{2,\theta_1^*,\psi^*}(x)^\mathsf{T}, \nabla_{\psi} f_{2,\theta_1^*,\psi^*}(x)^\mathsf{T}\right)^\mathsf{T}.$$

Then for all $\psi \in \mathcal{S}$ with $\psi \neq 0$, if $\eta_0 \in \mathbb{R}$, $\eta_1 \in \mathbb{R}^{d_1}$

$$\eta_0(f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}) + \eta_1^\mathsf{T} \nabla_{\theta_1} f_{1,\theta_1^*} = 0 \quad \Leftrightarrow \eta_1 = 0, \eta_2 = 0$$

Consistency

theorem

Given the mixture $f_{\theta_1,\psi,\alpha} = \alpha f_{1,\theta_1} + (1-\alpha) f_{2,\theta_1,\psi}$ and a sample $\mathbf{x}^n = (x_1, \cdots, x_n)$ issued from f_{1,θ_1^*}, under assumptions $B1 - B3$, and an $M > 0$ such that

$$\pi\left[(\alpha,\theta); \|f_{\theta,\alpha} - f^*\|_1 > M\sqrt{\log n/n}|\mathbf{x}^n\right] = o_p(1).$$

If $\alpha \sim \mathcal{B}(a_1, a_2)$, with $a_2 < d_2$, and if the prior $\pi_{\theta_1,\psi}$ is absolutely continuous with positive and continuous density at $(\theta_1^*, 0)$, then for $M_n \longrightarrow \infty$

$$\pi\left[|\alpha - \alpha^*| > M_n(\log n)^\gamma/\sqrt{n}|\mathbf{x}^n\right] = o_p(1), \quad \gamma = \max((d_1 + a_2)/(d_2 - a_2), 1)/2,$$

M-open case

When the true model behind the data is neither of the tested models, what happens?

- issue mostly bypassed by classical Bayesian procedures
- theoretically produces an α^* away from both 0 and 1
- possible (recommended?) inclusion of a Bayesian non-parametric model within alternatives

M-open case

When the true model behind the data is neither of the tested models, what happens?

- issue mostly bypassed by classical Bayesian procedures
- theoretically produces an α^* away from both 0 and 1
- possible (recommended?) inclusion of a Bayesian non-parametric model within alternatives

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