The vexing dilemma of Bayes tests of hypotheses and the predicted demise of the Bayes factor

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

Outline

Summary

Significance tests

Noninformative solutions

Jeffreys-Lindley paradox

Testing via mixtures

Testing issues

Hypothesis testing

- central problem of statistical inference
- 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)]

Testing 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

Testing 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

Some difficulties

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

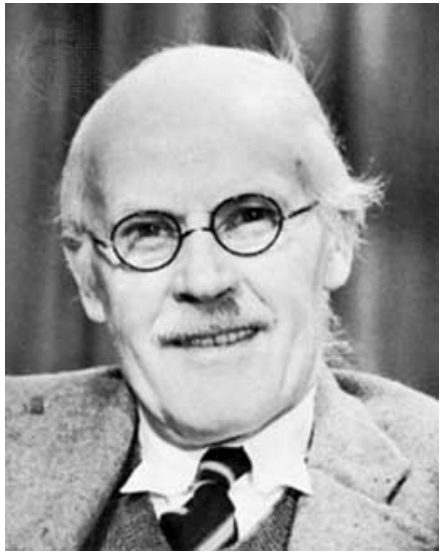
Some further difficulties

- long-lasting impact of the prior modeling, meaning the choice of the prior distributions on the parameter spaces of both models under comparison, despite overall consistency proof for Bayes factor
- discontinuity in use of improper priors since they are not justified in most testing situations, leading to many alternative if *ad hoc* solutions, where data is either used twice or split in artificial ways
- binary (*accept* vs. *reject*) outcome more suited for immediate decision (if any) than for model evaluation, in connection with rudimentary loss function

Some additional difficulties

- related impossibility to ascertain simultaneous misfit or to detect presence of outliers
- no assessment of uncertainty associated with decision itself
- difficult computation of marginal likelihoods in most settings with further controversies about which algorithm to adopt
- strong dependence of posterior probabilities on conditioning statistics, which in turn undermines their validity for model assessment, as exhibited in ABC model choice
- 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



data related questions

Given a dataset

$$x_1, \ldots, x_n$$

is it possible to answer a question related with the mechanism producing this data?

[Answer: No!]

For instance, is $\mathbb{E}[X] > 0$? Or, is $x_{n+1} = 10^4$ possible?

[Under some assumptions...]

data related questions

Given a dataset

$$x_1, \ldots, x_n$$

is it possible to answer a question related with the mechanism producing this data?

[Answer: No!]

For instance, is $\mathbb{E}[X] > 0$? Or, is $x_{n+1} = 10^4$ possible?

[Under some assumptions...]

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)*

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

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

Generalisation

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

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

Generalisation

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

[...fairly arbitrary!]

regressive illustration

caterpillar dataset from Bayesian Essentials (2013): predicting density of caterpillar nests from 10 covariates

	Estimate	Post. Var.	log10(BF)
(Intercept)	10.8895	6.8229	2.1873 (****)
X1	-0.0044	2e-06	1.1571 (***)
X2	-0.0533	0.0003	0.6667 (**)
X3	0.0673	0.0072	-0.8585
X4	-1.2808	0.2316	0.4726 (*)
X5	0.2293	0.0079	0.3861 (*)
X6	-0.3532	1.7877	-0.9860
X7	-0.2351	0.7373	-0.9848
X8	0.1793	0.0408	-0.8223
X9	-1.2726	0.5449	-0.3461
X10	-0.4288	0.3934	-0.8949

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

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

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

Normal example

Testing whether the mean α of a normal observation $X \sim \mathcal{N}(\alpha, s^2)$ is zero:

$$P(H_0|x) \propto \exp\left(-\frac{x^2}{2s^2}\right)$$

$$P(H_0^c|x) \propto \int \exp\left(-\frac{(x-\alpha)^2}{2s^2}\right) f(\alpha)\mathrm{d}\alpha$$

regressive illustation

caterpillar dataset from Bayesian Essentials (2013): predicting density of caterpillar nests from 10 covariates

	Estimate	Post. Var.	log10(BF)
(Intercept)	10.8895	6.8229	2.1873 (****)
X1	-0.0044	2e-06	1.1571 (***)
X2	-0.0533	0.0003	0.6667 (**)
X3	0.0673	0.0072	-0.8585
X4	-1.2808	0.2316	0.4726 (*)
X5	0.2293	0.0079	0.3861 (*)
X6	-0.3532	1.7877	-0.9860
X7	-0.2351	0.7373	-0.9848
X8	0.1793	0.0408	-0.8223
X9	-1.2726	0.5449	-0.3461
X10	-0.4288	0.3934	-0.8949

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

Noninformative proposals

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]

Vague proper priors are not the solution

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

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

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

Vague proper priors are not the solution

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

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

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

Vague proper priors are not the solution

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

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

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

Learning from the sample

Definition (Learning sample)

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

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

Learning from the sample

Definition (Learning sample)

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

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

Pseudo-Bayes factors

Idea

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

- π_i improper but $\pi_i(\cdot|x_{[i]})$ proper
- and

$$\frac{\int f_i(x_{[n/i]}|\theta_i)\,\pi_i(\theta_i|x_{[i]})\mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\,\pi_j(\theta_j|x_{[i]})\mathrm{d}\theta_j}$$

  independent of normalizing constant
- Use remaining $x_{[n/i]}$ to run test as if $\pi_j(\theta_j|x_{[i]})$ is the true prior

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)

Unexpected problems!

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

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

[Berger & Pericchi, 1996]

Unexpected problems!

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

Jeffreys–Lindley paradox

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

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

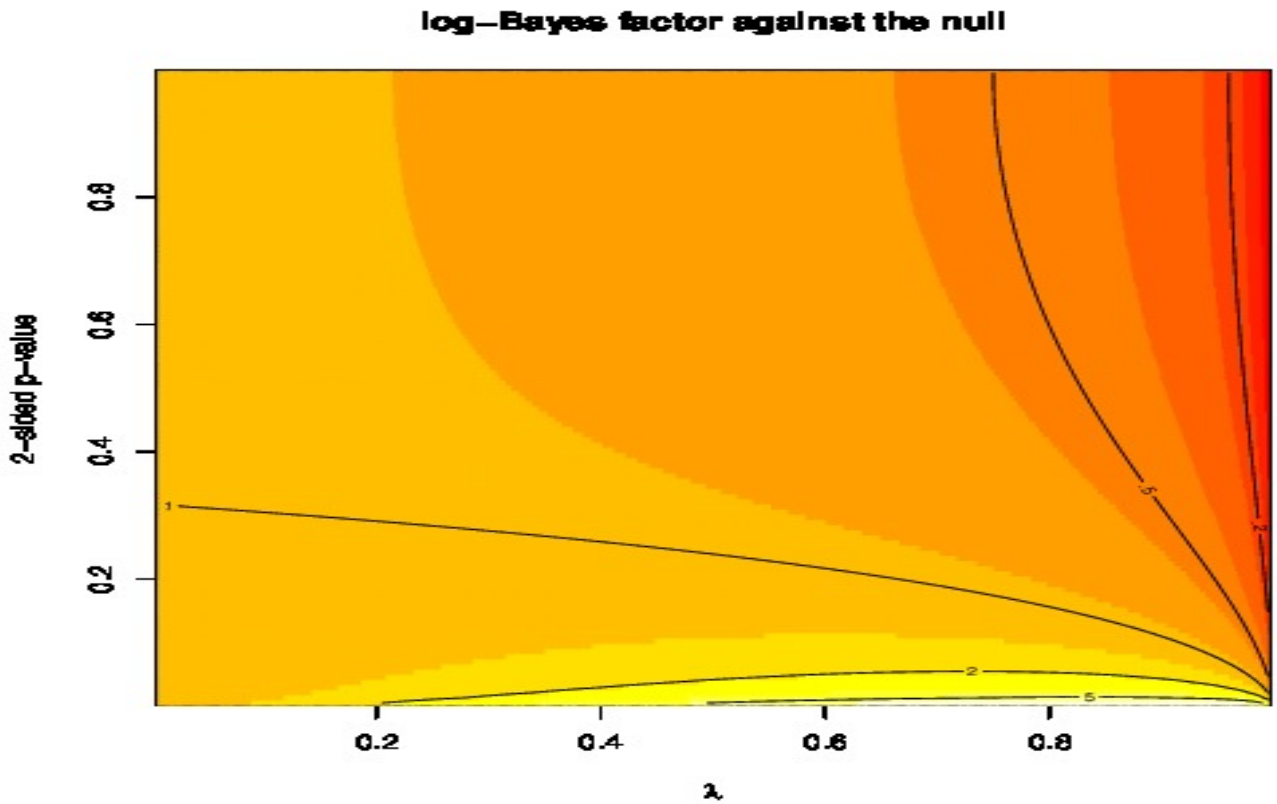
  for a given x, meaning H_0 is always accepted

[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]

Changing the testing perspective

Summary

Significance tests

Noninformative solutions

Jeffreys-Lindley paradox

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 the posterior distribution of the weight of a model, while moving from the artificial notion of the 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 an 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
- 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

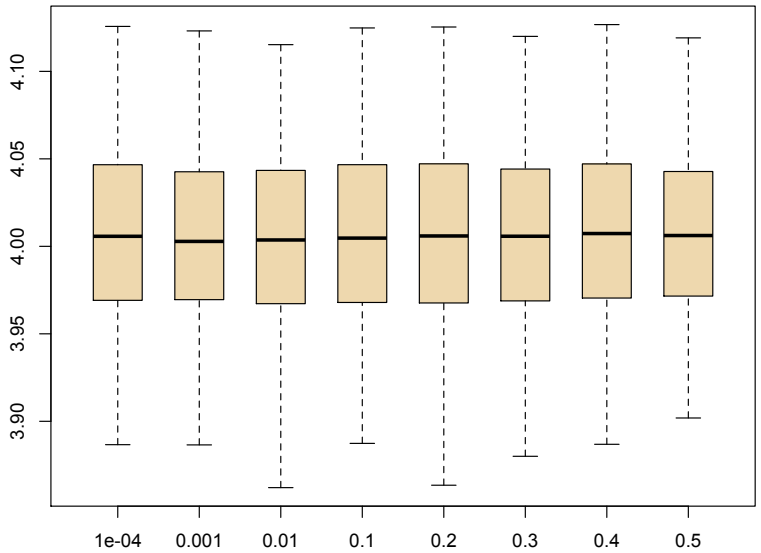
$$\alpha \sim \mathcal{B}e(n_1(\zeta) + a_0, n_2(\zeta) + a_0)$$

Exact Bayes factor opposing Poisson and Geometric

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

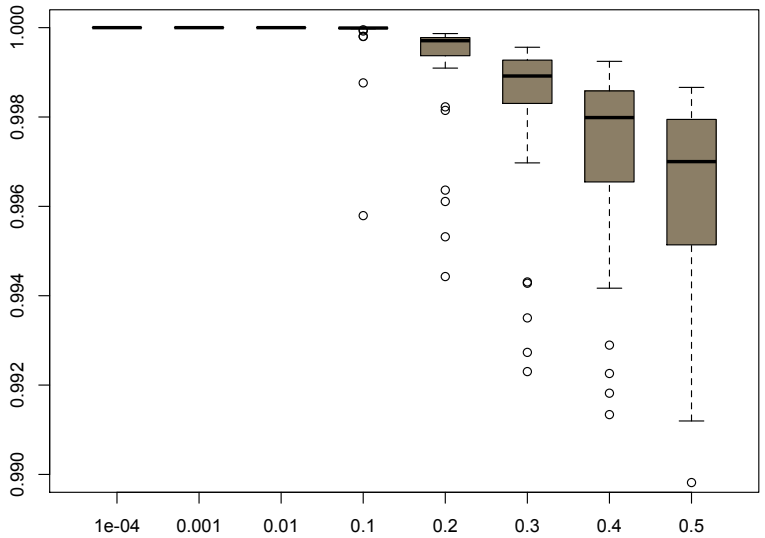
although undefined from a purely mathematical viewpoint

Parameter estimation



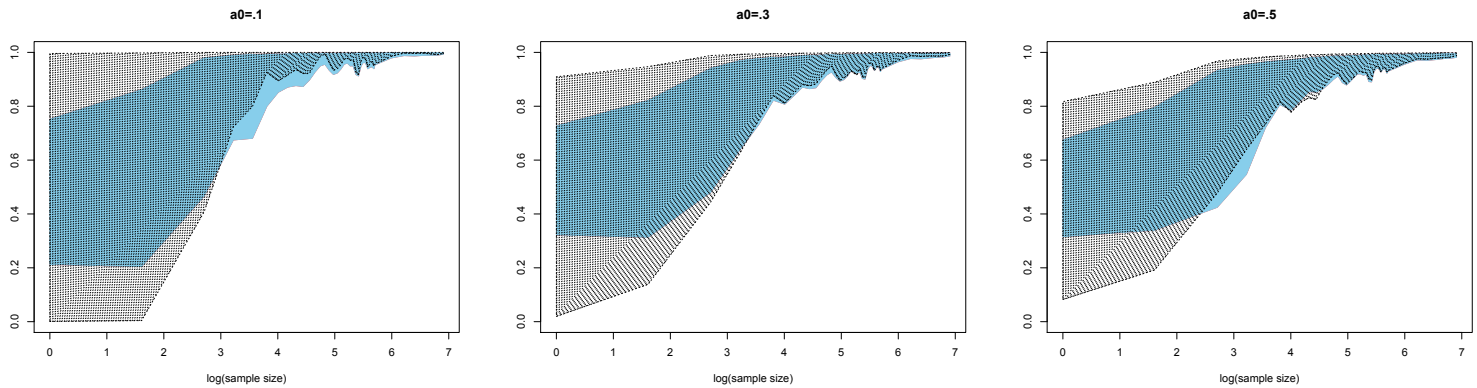
Posterior means of λ and medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Parameter estimation



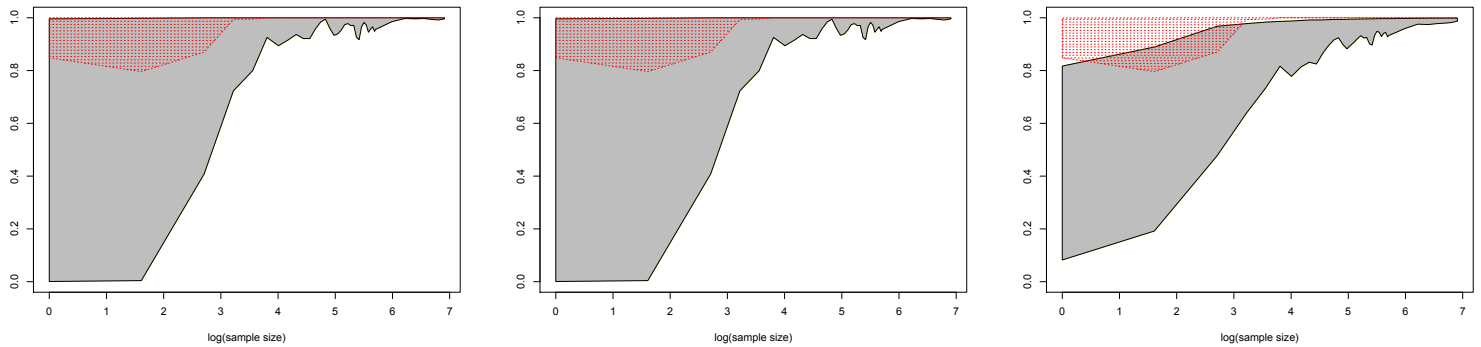
Posterior means of λ and medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$, for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

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

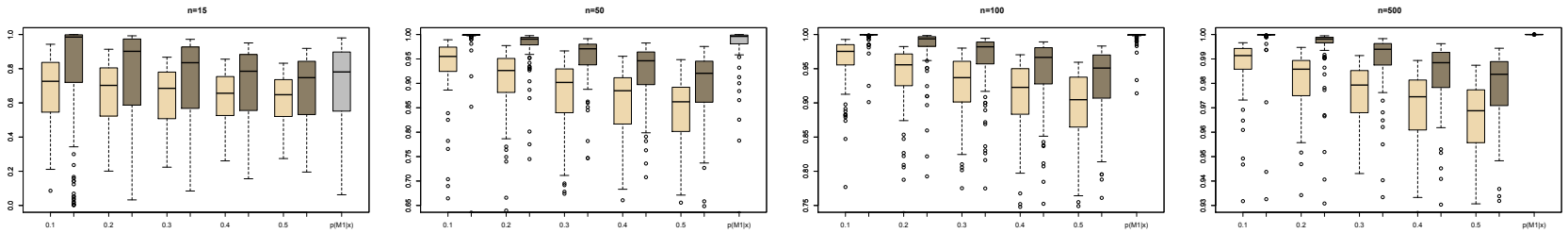
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

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

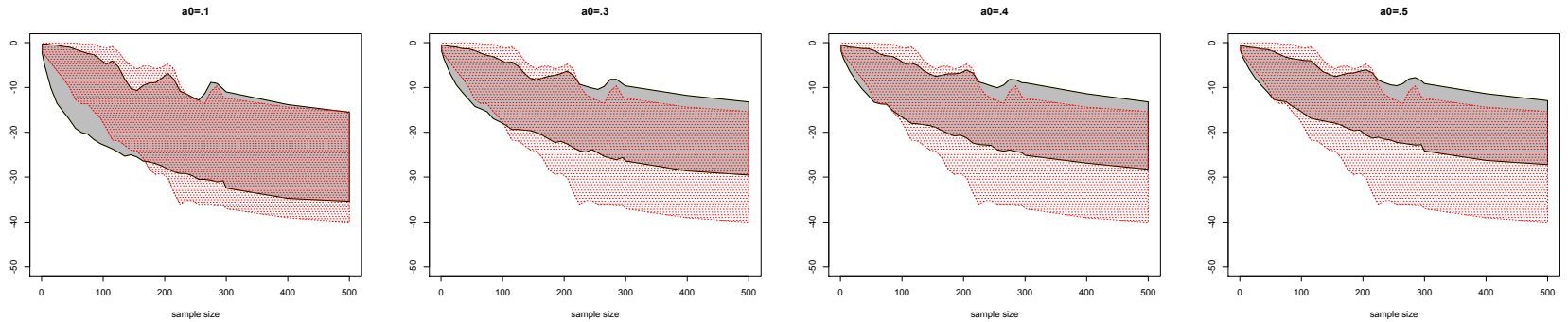
$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

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

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

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

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

Prior modelling

Under default g-prior

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

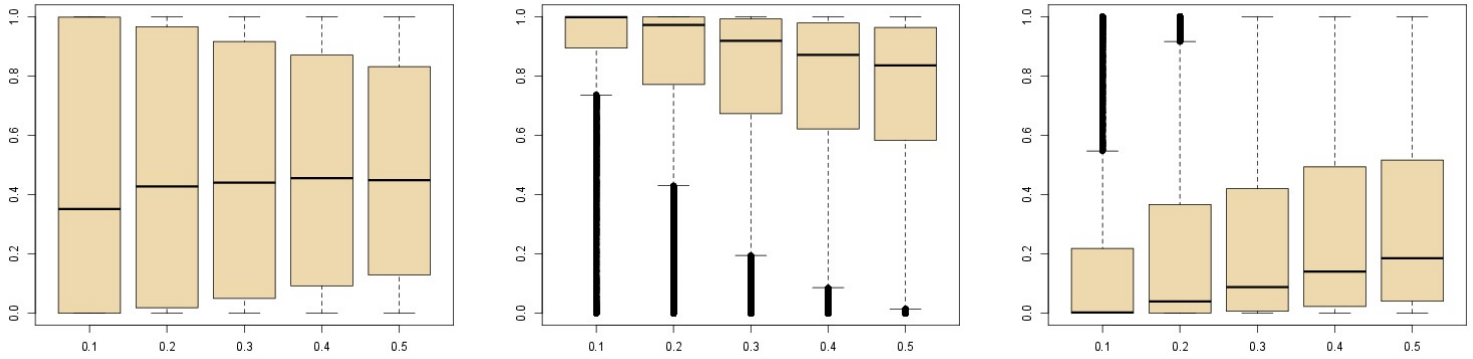
full conditional posterior distributions given allocations

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

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):

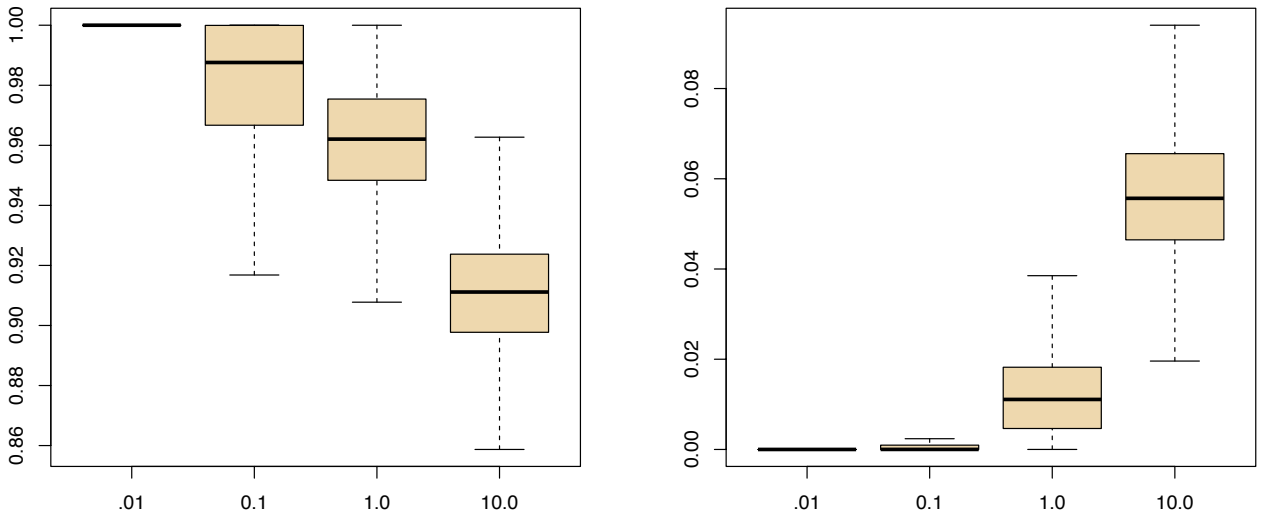
$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

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 $M_n \longrightarrow \infty$

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