Elaborating on Issues with Bayes Factors

Jorge N. Tendeiro   Henk A. L. Kiers

July 10, 2018

University of Groningen

Motivation

"The field of psychology is experiencing a crisis of confidence, as many researchers believe published results are not as well supported as claimed."[1]

Q: Why?

A: Among several other reasons (QRPs[2,3]), due to overreliance on NHST and p values.[4,5,6,7]

---

[1] Rouder (2014).
[2] John, Loewenstein, and Prelec (2012).
[3] Simmons, Nelson, and Simonsohn (2011).
[4] Edwards, Lindman, and Savage (1963).
[5] Cohen (1994).
[6] Nickerson (2000).
[7] Wagenmakers (2007).

Bayes factors are being increasingly advocated as a better alternative to NHST.[1,2,3,4,5]

We felt we did not know enough about Bayes factors (peculiarities, pitfalls, problems).

We surveyed the literature. Here we summarize what we found.

[1] Jeffreys (1961).
[2] Wagenmakers et al. (2010).
[3] Vampaemel (2010).
[4] Masson (2011).
[5] Dienes (2014).

Bayes factors: An X-ray

The Bayes factor[1,2] quantifies the change in prior odds to posterior odds due to the data observed.

- Two models to compare, for instance $\mathcal{M}_0 : \theta = 0$ vs $\mathcal{M}_1 : \theta \neq 0$.
- Data D.

By Bayes' rule ($i = 0, 1$):

$$p(\mathcal{M}_i|D) = \frac{p(\mathcal{M}_i)p(D|\mathcal{M}_i)}{p(\mathcal{M}_0)p(D|\mathcal{M}_0) + p(\mathcal{M}_1)p(D|\mathcal{M}_1)}.$$

Then

$$\underbrace{\frac{p(\mathcal{M}_0|D)}{p(\mathcal{M}_1|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}}_{\text{Bayes factor, } BF_{01}}.$$

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

- Typical interpretation, e.g., $BF_{01} = 5$:

  The data are five times more likely to have occurred under $\mathcal{M}_0$ than under $\mathcal{M}_1$.

- $BF_{01} \in [0, \infty)$:
  - $BF_{01} < 1 \longrightarrow$ Support for $\mathcal{M}_1$ over $\mathcal{M}_0$.
  - $BF_{01} = 1 \longrightarrow$ Equal support for either model.
  - $BF_{01} > 1 \longrightarrow$ Support for $\mathcal{M}_0$ over $\mathcal{M}_1$.

Bayes factor have been praised in many instances.[1,2,3,4,5]

Here we take a critical look at Bayes factors.

---

[1] Dienes (2011).
[2] Dienes (2014).
[3] Masson (2011).
[4] Vampaemel (2010).
[5] Wagenmakers et al. (2018).

1. Bayes factors are hard to compute. →
2. Bayes factors are sensitive to priors. →
3. Bayes factors are not posterior model probabilities. →
4. Bayes factors do not imply a model is correct. →
5. Interpretation of Bayes factors can be ambiguous. →
6. Bayes factors test model *classes*. →
7. Bayes factors $\longleftrightarrow$ parameter estimation. →
8. 'Default' Bayes factors lack justification. →
9. Bayes factors favor point $\mathcal{M}_0$. →
10. Bayes factors don't favor one-sided $\mathcal{M}_0$. →
11. Bayes factors favor $\mathcal{M}_a$. →
12. Bayes factors favor $\mathcal{M}_a$, II. →
13. Bayes factors may be problematic for nested models. →
14. Bayes factors and the replication crisis. →

Bayes factors are sensitive to priors

- Very well known.[1,2,3,4,5]
- Due to fact that the likelihood function is averaged over the prior to compute the marginal likelihood under a model.

Example: Bias of a coin[6]

- Three possible states: Two-headed, two-tailed, fair.
- $\mathcal{M}_0$: Two-headed *vs* $\mathcal{M}_1$: Not two-headed
- Data: Four heads out of four tosses.

Prior	p(heads) 0	.5	1	Intuition	BF_{01}	Lee & Wagenmakers (2014)
A	.01	.98	.01	Coin is fair	**16.2**	'Strong' evidence for $\mathcal{M}_0$
B	.33	.33	.33	Complete ignorance	**32**	'Very strong' evidence for $\mathcal{M}_0$
C	.49	.02	.49	Coin is unfair, either way	**408**	'Extreme' evidence for $\mathcal{M}_0$

The Bayes factors vary by as much as one order of magnitude.

[1] Kass (1993).
[2] Gallistel (2009).
[3] Vampaemel (2010).
[4] Robert (2016).
[5] Withers (2002).
[6] Lavine and Schervish (1999).

- The previous example is by no means unique or restricted to discrete random variables.[1,2]
- Varying priors may lead to results displaying support for different hypotheses.[3]
- Arbitrarily vague priors are not allowed because the null model would be invariably supported. So, in the Bayes Factor context, vague priors will predetermine the test result![4]
- However, counterintuitively, improper priors *might* work.[5]
- The problem cannot be solved by increasing sample size.[6,7,8]

This behavior of Bayes factors is in sharp contrast with estimation of posterior distributions.[9,10]

---

[1] Liu and Aitkin (2008).
[2] Berger and Pericchi (2001).
[3] Liu and Aitkin (2008).
[4] Morey and Rouder (2011).
[5] Berger and Pericchi (2001).
[6] Bayarri et al. (2012).
[7] Berger and Pericchi (2001).
[8] Kass and Raftery (1995).
[9] Gelman and Rubin (1995).
[10] Kass (1993).

How to best choose priors then?

- Some defend informative priors should be part of model setup and evaluation.[1]
- Other suggest using default/ reference/ objective, well chosen, priors.[2,3,4,5]
- Perform sensitivity analysis.

[1] Vampaemel (2010).
[2] Bayarri et al. (2012).
[3] Jeffreys (1961).
[4] Marden (2000).
[5] Rouder et al. (2009).

Bayes factors are not posterior model probabilities

Say that $BF_{01} = 32$; what does this mean?

After looking at the data, we revise our belief towards $\mathcal{M}_0$ by about 32 times.

Q: What does this imply concerning the probability of each model, given the observed data?
A: On its own, nothing at all!

Bayes factors are the multiplicative factor converting prior odds to posterior odds. They say nothing directly about model probabilities.

$$\underbrace{\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}}_{\text{prior odds}} \times \underbrace{\frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{M}_0|D)}{p(\mathcal{M}_1|D)}}_{\text{posterior odds}}$$

- Bayes factors say nothing about the plausability of each model in light of the data, that is, of $p(\mathcal{M}_i|D)$.
- Thus, Bayes factors $=$ rate of change of belief, not belief itself.[1]
- To compute $p(\mathcal{M}_i|D)$, prior model probabilities are needed:

$$p(\mathcal{M}_0|D) = \frac{\text{Prior odds} \times BF_{01}}{1 + \text{Prior odds} \times BF_{01}}, \quad p(\mathcal{M}_1|D) = 1 - p(\mathcal{M}_0|D).$$

Example

- Anna: Equal prior belief for either model.
- Ben: Strong prior belief for $\mathcal{M}_1$.
- $BF_{01} = 32$: Applies to Anna and Ben equally.

	$p(\mathcal{M}_0)$	$p(\mathcal{M}_1)$	BF_{01}	$p(\mathcal{M}_0\|D)$	$p(\mathcal{M}_1\|D)$	Conclusion
Anna	.50	.50	32	**.970**	.030	Favors $\mathcal{M}_0$
Ben	.01	.99		.244	**.756**	Favors $\mathcal{M}_1$

[1] Edwards, Lindman, and Savage (1963).

Bayes factors $\longleftrightarrow$ parameter estimation

- Frequentist two-sided significance tests and confidence intervals (CIs) are directly related:
  The null hypothesis is rejected iff the null point is outside the CI.
- This is not valid in the Bayesian framework.[1]

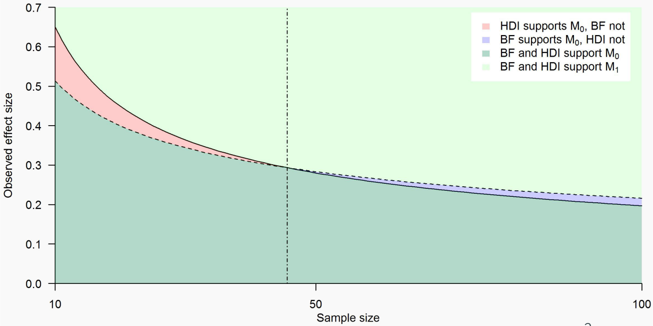

Figure 1: Data: $Y_i \sim N(\mu, \sigma)$. $\mathcal{M}_0 : \delta = 0$ vs $\mathcal{M}_1 : \delta \sim N(0, \sigma_0^2)$, $\delta = \mu/\sigma$.

[1] Kruschke and Liddell (2018b).

Bayes factors favor point $\mathcal{M}_0$

- NHST is strongly biased against the point null model $\mathcal{M}_0$.[1,2,3,4]
- In other words, $p(\mathcal{M}_0|D)$ and p values do not agree.
  (Yes, they are conceptually different![5])
- The discrepancy worsens as the sample size increases.

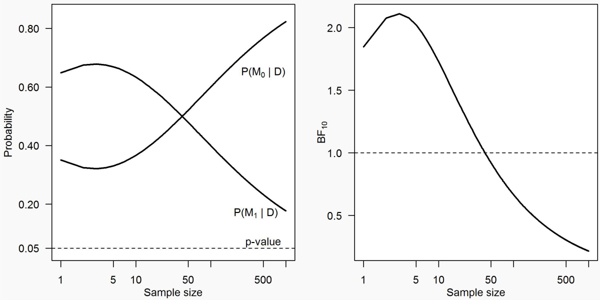

Figure 2: Data: $Y_i \sim N(\mu, 1)$. $\mathcal{M}_0 : \mu = 0$ *vs* $\mathcal{M}_1 : \mu \sim N(0, 1)$.

---

[1] Edwards, Lindman, and Savage (1963).
[2] Dickey (1977).
[3] Berger and Sellke (1987).
[4] Sellke, Bayarri, and Berger (2001).
[5] Gigerenzer (2018).

- In this example, for $n > 42$ one rejects $\mathcal{M}_0$ under NHST whereas $BF_{10} < 1$ (indicating support for $\mathcal{M}_0$).
- In sum: Bigger ESs are needed for Bayes factor to sway towards $\mathcal{M}_1$. But, how much bigger?

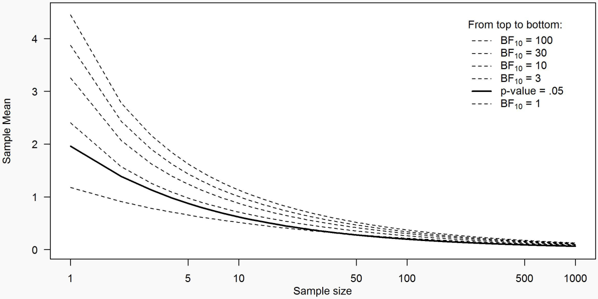

Figure 3: ESs required by BF_{10}, based of Jeffreys (1961) taxonomy.

Calibrate Bayes factors $\longleftrightarrow$ p values?[1,2]

[1] Wetzels et al. (2011).

[2] Jeon and De Boeck (2017).

Bayes factors don't favor one-sided $\mathcal{M}_0$

- Surprisingly, the previous result does not hold for one-sided $\mathcal{M}_0$ (e.g., $\mathcal{M}_0 : \mu < 0$).[1,2]
- In this case, $p(\mathcal{M}_0|D)$ and p values can be very close under a wide range of priors.

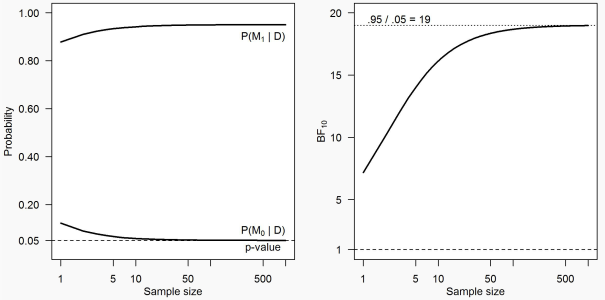

Figure 4: Data: $Y_i \sim N(\mu, 1)$. $\mathcal{M}_0 : \mu \sim N^+(0, 1)$ vs $\mathcal{M}_1 : \mu \sim N^-(0, 1)$.

---

[1] Pratt (1965).

[2] Casella and Berger (1987).

Tuning just-significant ESs with Bayes factors:

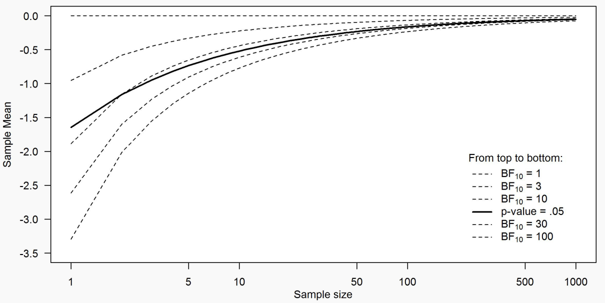

Figure 5: ESs required by BF_{10}, based of Jeffreys (1961) taxonomy.

- $p(\mathcal{M}_0|D)$ can be equal or even smaller than the p value.[1]
- 'p values overstate evidence against $\mathcal{M}_0$' $\longrightarrow$ Not always.[2]

Who to blame for this state of affairs?

We suggest the nature of the point null hypothesis; we are not alone.[3,4]
But others have argued in favor point of null hypotheses.[5,6,7,8,9,10]

'True' point hypotheses, really?![11,12,13]

---

[1] Casella and Berger (1987).
[2] Jeffreys (1961).
[3] Casella and Berger (1987).
[4] Vardeman (1987).
[5] Berger and Delampady (1987).
[6] Kass and Raftery (1995).
[7] Gallistel (2009).
[8] Konijn et al. (2015).
[9] Marden (2000).
[10] Morey and Rouder (2011).
[11] Berger and Delampady (1987).
[12] Cohen (1994).
[13] Morey and Rouder (2011).

Bayes factors favor $\mathcal{M}_a$

- Unless $\mathcal{M}_0$ is exactly true, $n \to \infty \implies BF_{01} \to 0$.
- Thus, both BF_{01} and the p value approach 0 as n increases.
- It has be argued that this is a good property of Bayes factors (they are information consistent).[1]
- However, BF_{01} does ignore 'practical significance', or magnitude of ESs.[2]
- Meehl's paradox: For true negligible non-zero ESs, data accumulation should make it easier to reject a theory, not confirm it.[3,4]

---

[1] Ly, Verhagen, and Wagenmakers (2016).
[2] Morey and Rouder (2011).
[3] Meehl (1967).
[4] Kruschke and Liddell (2018b).

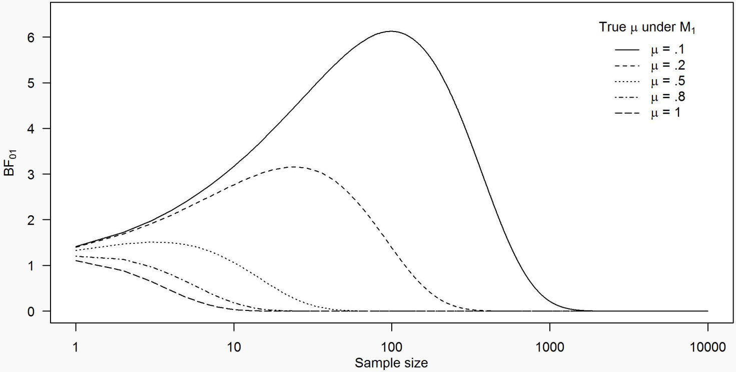

Figure 6: Data: $Y_i \sim N(\mu, 1)$. $\mathcal{M}_0 : \mu = 0$ *vs* $\mathcal{M}_1 : \mu \sim N(0, 1)$.

Bayes factors favor $\mathcal{M}_a$, II

- Consider $\mathcal{M}_0 : \theta = \theta_0$ *vs* $\mathcal{M}_0 : \theta \neq \theta_0$.
- As $n \to \infty$, Bayes factors accumulate evidence in favor of true $\mathcal{M}_1$ much faster than they accumulate evidence in favor of true $\mathcal{M}_0$.
- I.e., although Bayes factors allow drawing support for either model, they do so asymmetrically.[1]

[1] Johnson and Rossell (2010).

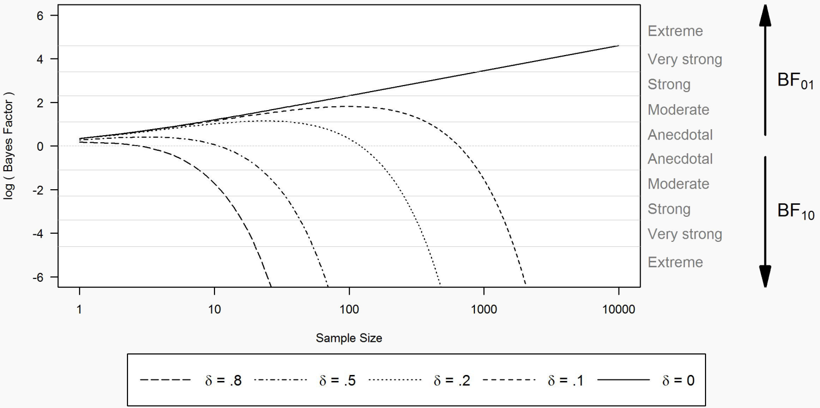

Figure 7: Data: $Y_i \sim N(\mu, \sigma)$. $\mathcal{M}_0 : \delta = 0$ vs $\mathcal{M}_1 : \delta \sim N(0, \sigma_0^2)$, $\delta = \mu/\sigma$.

Bayes factors and the replication crisis

- It is increasingly difficult to ignore the current crisis of confidence in psychological research.
- Several key papers and reports made the ongoing state of affairs unbearable.[1,2,3,4,5,6]
- Some attempts to mitigate the problem have been put forward, including pre-registration and recalibration.[7,8]
- Some have suggested that a shift towards Bayesian testing is welcome.[9,10,11]

Would Bayes factors contribute to improving things?

[1] Ioannidis (2005).
[2] Simmons, Nelson, and Simonsohn (2011).
[3] Bem (2011).
[4] Wicherts, Bakker, and Molenaar (2011).
[5] John, Loewenstein, and Prelec (2012).
[6] Open Science Collaboration (2015).
[7] Benjamin et al. (2018).
[8] Lakens et al. (2018).
[9] Vampaemel (2010).
[10] Konijn et al. (2015).
[11] Dienes (2016).

What Bayes factors promise to offer might not be what researchers and journals are willing to use.[1]

- It has not yet been shown that the Bayes factors' ability to draw support for $\mathcal{M}_0$ will alleviate the bias against publishing null results ("lack of effects" are still too unpopular).
  Bayes factors need not be aligned with current publication guidelines.
- 'B-hacking'[2] is still entirely possible. New QRPs lurking around the corner?

[1] Savalei and Dunn (2015).

[2] Konijn et al. (2015).

Now what?

We think that:

- The use, abuse, and misuse of NHST and p values are problematic. The statistical community is aware of this.[1]
- Bayes factors are an interesting alternative, but they do have limitations of their own.
- In particular, Bayes factors are also based on 'dichotomous modeling thinking': Given two models, which one is to be preferred?
  We favor a more holistic approach to model comparison.
- Bayes factors provide no direct information concerning effect sizes, their magnitude and uncertainty.[2,3] This is sorely missed by this approach.

[1] Wasserstein and Lazar (2016). [2] Wilkinson (1999). [3] Kruschke and Liddell (2018a).

What to do?

- Truly consider whether testing is what you need.
- In particular, point hypotheses seem prone to trouble.
  How realistic are these hypotheses?
- Do estimation![1,2,3]
  Perform inference based on the entire posterior distribution.
  Report credible values. Compute posterior probabilities.

[1] Cohen (1994).
[2] Kruschke (2011).
[3] van der Linden and Chryst (2017).

There are other tools, also based on the Bayesian paradigm, worth considering. These include:

- Bayes model averaging.[1]
- Generalization criterion.[2]
- Deviance information criterion.[3]
- Mixture model estimation.[4,5]
- Posterior predictive loss.[6]
- Posterior likelihood ratio.[7]
- Posterior predictive methods.[8,9,10]

---

[1] Hoeting et al. (1999).
[2] Liu and Aitkin (2008).
[3] Spiegelhalter et al. (2002).
[4] Kamary et al. (2014).
[5] Robert (2016).
[6] Gelfand and Ghosh (1998).
[7] Aitkin, Boys, and Chadwick (2005).
[8] Vehtari and Lampinen (2002).
[9] Vehtari and Ojanen (2012).
[10] Gelman et al. (2013).

Thank you

j.n.tendeiro@rug.nl

Bayes factors are hard to compute

Bayes factors are hard to compute

$$BF_{01} = \frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_1)}.$$

Bayes factors are ratios of marginal likelihoods:

$$P(D|\mathcal{M}_i) = \int_{\Theta_i} p(D|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta$$

- The marginal likelihoods, $P(D|\mathcal{M}_i)$, are hard to compute in general.
- Resort to (not straightforward) numerical procedures[1,2]
- Alternatively, use software with prepackaged default priors and data models[3,4] (limited to specific models).

---

[1] Chen, Shao, and Ibrahim (2000).
[2] Gamerman and Lopes (2006).
[3] JASP Team (2018).
[4] Morey and Rouder (2018).

Bayes factors do not imply a model is correct

Bayes factors do not imply a model is correct

- A large Bayes factor, say, $BF_{10} = 100$, may mislead one to belief that $\mathcal{M}_1$ is true or at least more useful.
- Bayes factors are only a measure of relative plausibility among two competing models.
- $\mathcal{M}_1$ might actually be a dreadful model (e.g., lead to horribly wrong predictions), but simply less dreadful than its alternative $\mathcal{M}_0$.[1]
- Bayes factors provide no absolute evidence supporting either model under comparison.[2]
- Little is known as to how Bayes factors behave under model misspecification (but see[3]).

[1] Rouder (2014).

[2] Gelman and Rubin (1995).

[3] Ly, Verhagen, and Wagenmakers (2016).

Interpretation of Bayes factors can be ambiguous

Interpretation of Bayes factors can be ambiguous

- Bayes factors are a continuous measure of evidence in $[0, \infty)$:
  - $BF_{01} > 1$: Data are more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$. The larger BF_{01}, the stronger the evidence for $\mathcal{M}_0$ over $\mathcal{M}_1$.
  - $BF_{01} < 1$: Data are more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$. The smaller BF_{01}, the stronger the evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$.
- But, how 'much more' likely?
- Answer is not unique: Qualitative interpretations of strength are subjective (what is weak?, moderate?, strong?).[1,2,3,4]

This is not a problem of Bayes factor per se, but of practitioners requiring qualitative labels for test results.

---

[1] Jeffreys (1961).

[2] Kass and Raftery (1995).

[3] Lee and Wagenmakers (2013).

[4] Dienes (2016).

Bayes factors test model *classes*

Bayes factors test model *classes*

Consider testing $\mathcal{M}_0 : \theta = \theta_0$ vs $\mathcal{M}_1 : \theta \neq \theta_0$. Then

$$B_{01} = \frac{p(D|\mathcal{M}_0)}{p(D|\mathcal{M}_1)}, \quad \text{with} \quad p(D|\mathcal{M}_1) = \int p(D|\theta, \mathcal{M}_1) p(\theta|\mathcal{M}_1) d\theta.$$

- $p(D|\mathcal{M}_1)$ is a weighted likelihood for a model class:
  Each parameter value θ defines one particular model in the class.
- Bayes factors as ratios of likelihoods of model classes.[1]
- E.g., $BF_{01} = 1/5$: The data are five times more likely under the model class under $\mathcal{M}_1$, averaged over its prior distribution, than under $\mathcal{M}_0$.
- Catch: *The most likely model class need not include the true model that generated the data.*
  I.e., the Bayes factor may fail to indicate the class that includes the data-generating model (in case it exists, of course).[2]

[1] Liu and Aitkin (2008).

[2] Liu and Aitkin (ibid.).

‘Default’ Bayes factors lack justification

'Default' Bayes factors lack justification

- Priors matter a lot for Bayes factors.
- 'Objective' bayesians advocate using predefined priors for testing.[1,2,3]
- Albeit convenient, default priors lack empirical justification.[4]
- 'Objective priors' were derived under strong requirements[5,6] , which impose strong restrictions on the priors ("appearance of objectivity"[7]).
- Defaults are only useful to the extent that they adequately translate one's beliefs.[8,9]
- Some default priors, like the now famous JZS prior[10,11,12] , still require a specification of a scale parameter. Its default value has also changed over time.[13,14]

[1] Jeffreys (1961).
[2] Berger and Pericchi (2001).
[3] Rouder et al. (2009).
[4] Robert (2016).
[5] Bayarri et al. (2012).
[6] Berger and Pericchi (2001).
[7] Berger and Pericchi (ibid.).
[8] Kruschke (2011).
[9] Kruschke and Liddell (2018a).
[10] Jeffreys (1961).
[11] Zellner and Siow (1980).
[12] Rouder et al. (2009).
[13] Rouder et al. (ibid.).
[14] Morey and Rouder (2018).

Bayes factors may be problematic for nested models

Bayes factors may be problematic for nested models

- $\mathcal{M}_0$ is nested in $\mathcal{M}_1$ when $\mathcal{M}_0$ is a constrained form of $\mathcal{M}_1$. Example:

$$\mathcal{M}_0 : \theta = \theta_0 \quad \textit{vs} \quad \mathcal{M}_1 : \theta \neq \theta_0.$$

- Bayes factors were originally developed for nested models.[1]
- To compute BF_{01}, all parameters other than θ must be integrated out from both models. These are referred to as common or nuisance parameters.
- Vague priors over 'common' parameters are suggested to work (!!).[2]
- Usual strategy used by default Bayes factors:
  Use the same prior for the 'common' parameters under both models.

[1] Jeffreys (1939).

[2] Kass and Raftery (1995).

Bayes factors may be problematic for nested models

Problem
Distributional properties of the common parameters may change between models.[1,2]

Example
SD of residuals in nested regression models.

These are, more appropriately, "approximately common parameters".[3]

[1] Berger and Pericchi (2001).

[2] Robert (2016).

[3] Sinharay and Stern (2002).