Rejoinder

Harlan Campbell[*] and Paul Gustafson[†]

Introduction

We are grateful to all the discussants for their thoughtful perspectives. Specifically, Guha and Rizzelli offer compelling mathematical arguments that help us better understand when a frequentist's confidence interval and a Bayesian's credible interval will agree. Lad, and Held and Pawel consider, each in their own insightful way, the predictive probability distribution of a future observation. Srakar suggests that one might avoid the "strange behaviour" of the spike and slab prior by considering a peri-null prior instead. Clarke covers a lot of ground, ranging from concrete extensions beyond the canonical one-parameter problem, to more conceptual musings on hypothesis testing writ large. Rice provides much food for thought about being more explicitly decision-theoretic and considering different loss functions. Ly, Boehm, and Grünwald point to desirable properties of the anytime-valid confidence sequence based on inverting the Bayes factor decision, analogous to inverting a frequentist test to get a confidence interval. Finally, Heck, and Johnson and Datta offer perhaps the strongest opinions about the merits of the spike and slab prior and the relationship between estimation and testing.

When it comes to testing, the procedure of rejecting a point-null hypothesis (e.g., $H_0 : \theta = 0$) if and only if the null value is outside of the posterior credible interval ensures compatibility between hypothesis testing and parameter estimation. Those who reject a point-null hypothesis whenever the Bayes factor (BF) is above a certain threshold are also following this procedure with a (perhaps implicitly specified) spike and slab prior (Campbell and Gustafson, 2023). Evidently, not all believe that this compatibility is desirable. Heck, for instance, writes in favour of a "conceptual separation between hypothesis testing of a point-null hypothesis and parameter estimation," and suggests that "the estimate for the test parameter θ is interesting only if the data provide sufficient evidence for the effect."

In our opinion, agreement between testing and estimation is fundamentally appealing. We see the ongoing angst about the role of p-values (see Matthews, 2021 for a recent commentary) as partly fueled by a desire to think of testing and estimation separately. As well, the substantial, and often tricky, literature on mitigating challenges of "inference after selection" (see Kuchibhotla et al., 2022 for a recent review) arises from wanting said separation. So instead of debating the merits of compatibility between testing and estimation, we focus this rejoinder on the question of whether or not specifying a spike and slab prior (either implicitly with a BF, or explicitly) is ever a good idea.

[*]Department of Statistics, University of British Columbia; Health Economics & Outcomes Research, Precision AQ, harlan.campbell@stat.ubc.ca

[†]Department of Statistics, University of British Columbia, gustaf@stat.ubc.ca

Is a spike and slab prior ever a good idea?

Ideally, one should specify a prior distribution for θ that aligns with the true parameter-generating distribution (PGD), though before saying more, it is worth considering what the PGD (true or otherwise) actually is. One take, emphasized by Gustafson and Greenland (2009), is to think of the PGD as describing the true parameter values for a series of *different* scientific relationships that a "lab" will study over time. While still somewhat amorphous, this is more concrete than simply saying the PGD is what "Mother Nature" draws from to set the state of the world before data are generated. However, regardless of the narrative employed, the math is clear. If the prior employed matches the PGD, posterior credible intervals are guaranteed to have certain desirable decision-theoretic properties as well as correct "labwise coverage" (i.e., taken with respect to the distribution of θ and the data jointly, the probability that θ is within the bounds of the posterior $(1-\alpha)\%$ credible interval will be equal to $(1-\alpha)$). However, even upon adopting the "lab" interpretation of the PGD, without omniscient powers the distribution cannot be known exactly. Thus, approximations, as well as practical considerations, must inform one's prior specification.

Based on the notation that "the point-null is never true" (Nester, 1996; Gelman and Carlin, 2017; Gelman et al., 2013), many might argue that the spike and slab prior is never a reasonable approximation of the PGD. However, there are scenarios in which a spike and slab prior seems appropriate. For instance, spike and slab priors may be well suited for "semi-continuous" parameters (e.g., a parameter corresponding to rainfall (equal to exactly zero if it did not rain, positive and continuous otherwise), or to the financial gain/loss from a possible investment (equal to exactly zero if the investment was never made, continuous otherwise)). When θ corresponds to a treatment effect, it may be more difficult (although not impossible) to justify a spike and slab prior. Gelman et al. (2013) for example, are unequivocal on this matter, writing: "we do not like models that assign a positive probability to the event $\theta = 0$, if θ is some continuous parameter such as a treatment effect." However, even if the spike and slab prior is not a realistic approximation of the PGD, it might still be a good pragmatic choice.

Advantages

Aside from providing some computational conveniences (e.g., some high-dimensional models can be fit more easily with spike and slab priors (Castillo et al., 2015)), the spike and slab prior offers one clear advantage: the possibility of a seemingly objective procedure to "accept the null hypothesis." The procedure is as follows. One accepts the null if and only if $\Pr(\theta = 0|data) \geq (1-\alpha)$. This procedure is often described in terms of the Bayes factor whereby one accepts the null whenever $\text{BF} \leq (1-\alpha)/\alpha$, with implied prior model odds of 1:1; see Lavine and Schervish (1999); Campbell and Gustafson (2023).

Without a spike at zero in the prior, the posterior probability on the point null will necessarily be equal to zero (i.e., $\Pr(\theta = 0|data) = 0$, regardless of the data), and one can only "accept the null" approximately. Typically this is done by defining an "equivalence margin", Δ, and evaluating whether $\Pr(\theta \in \Delta|data) \geq (1-\alpha)$. Regrettably, defining Δ

(also known as a "region of practical equivalence") requires some degree of subjectivity about what values of θ can be considered negligible; see Campbell and Gustafson (2021, 2024); Schwaferts and Augustin (2020).

The procedure of specifying a spike and slab prior and "accepting the null" whenever $\Pr(\theta = 0|data) \geq (1-\alpha)$ certainly has appeal as being the natural opposite to rejecting the null whenever the point value is outside of the posterior credible interval. However, in practice, we see two drawbacks.

First, small, seemingly inconsequential changes in how the slab portion of the spike and slab prior is defined can have large impacts on the posterior; this is well documented, typically being framed as the BF being highly sensitive to the prior variance under the alternative model (e.g., Gelman et al., 2013). Careful sensitivity analyses and/or the use of so-called "default" or "intrinsic" priors may provide some remedy (Rouder and Morey, 2012; Womack et al., 2014), but the issue remains delicate.

Second, as Johnson and Datta explain, the slow rate of convergence makes it difficult to accept the null unless sample sizes are very large. Johnson and Datta recommend using a "non-local" (NL) density for the slab part of the prior so that it is easier to distinguish between the null and alternative hypotheses when the null is true. This solution however, requires one to make a rather subjective choice with respect to how exactly to define the NL density.

To illustrate, suppose a NL normal moment density with modes at ± 0.3 is used to define the slab part of a spike and slab prior on θ, as in Johnson and Datta's Figure 1. Then with data consisting of $Z = 1.645$ and $n = 1{,}000$, one would accept the null since $\Pr(\theta = 0|data) = 0.958$ (or equivalently BF $= 1/22.8$). (Note that a 95.8%CrI consists solely of $\{0\}$, while the strict 95.0%CrI is not defined.) If instead a NL normal moment density with modes at ± 0.1 is specified, one would *not* accept the null since $\Pr(\theta = 0|data) = 0.594$ (or equivalently, BF $= 1/1.5$). In this case, the equal-tailed 95%CrI does not exist but a lopsided interval of [0.000, 0.100) does have posterior probability mass of exactly 0.950, since $\Pr(\theta < 0|data) = 0.003$ and $\Pr(\theta < 0.100|data) = 0.953$. With a small difference in specifying the NL normal moment density (e.g., modes at ± 0.3 vs. ± 0.1) having a substantial impact on one's conclusions, effectively the need to subjectively determine what values of θ can be considered negligible remains.

Disadvantages

Aside from the minor inconvenience that certain credible intervals cannot be defined exactly, we see one main disadvantage of the spike and slab prior: the posterior credible interval obtained with a spike and slab prior will not converge asymptotically to the frequentist confidence interval. In other words, specifying a spike and slab prior allows for the possibility of a disagreement between the frequentist and Bayesian analyses even with very large sample sizes. This is the crux of the Jeffreys-Lindley paradox but the extent to which this is problematic is clearly a matter of opinion.

Srakar suggests that one might avoid this "strange behaviour" by considering a peri-null prior instead. Indeed, Cherry (2023) recently explains how if the point null is

replaced by a narrow distribution, the Bayesian and frequentist answers do not diverge (if one asks the same question with both approaches). However, a peri-null prior will require a rather subjective choice to define its narrowness and "seemingly inconsequential changes in prior specification may asymptotically yield fundamentally different results" (Ly and Wagenmakers, 2022).

Johnson and Datta suggest that, regardless of the potential for disagreement between Bayesian and frequentist answers, a credible interval based on a spike and slab prior is still desirable on the grounds of consistency, i.e., asymptotically a type I error rate of zero is achieved. (Johnson and Datta: "[T]here is α probability that the confidence interval will not contain $\{0\}$, no matter how large n is even when the null hypothesis is true. In this regard, there is a fundamental disagreement between the intervals. The credible interval [with a spike and slab prior] is consistent and the confidence interval is not.") A desire for this type of consistency is certainly understandable (and brings to mind the definition of "Chernoff consistency" given in Shao, 2008). For balance, however, note that other routes to consistency are possible. A $(1-\alpha)\%$ posterior credible interval based on a continuous prior (or a classical $(1-\alpha)\%$ confidence interval, for that matter) can easily be made consistent by letting $\alpha = \alpha_n \to 0$ as $n \to \infty$. (And for some, the explicit control of the rate at which the type 1 error rate vanishes might be appealing.)

Reporting

If one does use a spike and slab prior, certain credible intervals cannot be defined exactly and it is not immediately clear how to report their results. Several discussants suggest redefining the $(1-\alpha)\%$ credible interval conservatively as the interval just big enough to have support of at least $1-\alpha$. This seems like a reasonable approach. Alternatively, one could simply report the $(1-\gamma)\%$ credible interval where γ is the value closest to α for which the credible interval can be defined exactly. This follows the recommendation of Young et al. (2005) that is cited by Rice.

Heck suggests reporting the BF as well as the "default credible interval" (i.e., the credible interval obtained without the spike in the prior, what Johnson and Datta call the "conditional posterior credible interval under the alternative"). The BF (as it is most often interpreted) corresponds to a spike and slab prior, which is of course a very different prior distribution than the prior implied by reporting the "default credible interval". In principle, we are not against reporting a variety of different estimates arising from different priors. Our concern is that making these two things the default inferential summaries, without emphasizing the wildly different priors driving them, obscures what is really going on. We would therefore argue for emphatic labeling of what kind of prior drives what answer.

We also recommend that, at a minimum, if one reports a BF, then one should also report the corresponding posterior estimate based on the implied spike and slab prior; see Campbell and Gustafson (2023). This is rarely done. For example, in a recent Bayesian re-analysis of randomized trials data (Pittelkow et al., 2024), BFs are reported without any associated credible intervals at all. The authors conclude that their analyses with the BF are "more informative than traditional measures of uncertainty such as the

confidence interval, because it allowed us to disambiguate between absence of evidence and evidence of absence." Perhaps, but a BF without any accompanying credible interval nor any explanation/justification of the (implied) priors is also easily misinterpreted (Tendeiro et al., 2024; Campbell and Gustafson, 2024).

Perhaps something all can agree upon is the usefulness of reporting estimates and credible intervals for a hypothetical future observation, Y_{n+1}. Predicting the "next observation", while perhaps not the main scientific objective, certainly helps one understand the true degree of posterior uncertainty. It is also reassuring that, as discussed by Lad, and Held and Pawel, the prior on θ will have less impact on the credible interval for Y_{n+1}. Moreover, as demonstrated by Held and Pawel, the confidence interval and credible interval for Y_{n+1} will converge, even with a spike and slab prior on θ (but the rate of convergence will be much slower).

References

Campbell, H. and Gustafson, P. (2021). "What to make of equivalence testing with a post-specified margin?" *Meta-Psychology*, 5. 981

Campbell, H. and Gustafson, P. (2023). "Bayes factors and posterior estimation: Two sides of the very same coin." *The American Statistician*, 77(3): 248–258. MR4621941. doi: https://doi.org/10.1080/00031305.2022.2139293. 979, 980, 982

Campbell, H. and Gustafson, P. (2024). "The Bayes factor, HDI-ROPE, and frequentist equivalence tests can all be reverse engineered—almost exactly—from one another: Reply to Linde et al. (2021)." *Psychological Methods*. 981, 983

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 980

Cherry, J. L. (2023). "Comment on "History and nature of the Jeffreys–Lindley paradox" by J. L. Cherry (Section 2 of "The Jeffreys–Lindley paradox: an exchange")." *Archive for History of Exact Sciences*, 77(4): 443–449. MR4604374. doi: https://doi.org/10.1007/s00407-023-00310-4. 981

Gelman, A. and Carlin, J. (2017). "Some natural solutions to the p-value communication problem—and why they won't work." *Journal of the American Statistical Association*, 112(519): 899–901. MR3735346. doi: https://doi.org/10.1080/01621459.2017.1311263. 980

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*, third edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. URL https://books.google.ca/books?id=ZXL6AQAAQBAJ MR3235677. 980, 981

Gustafson, P. and Greenland, S. (2009). "Interval estimation for messy observational data." *Statistical Science*, 328–342. MR2757434. doi: https://doi.org/10.1214/09-STS305. 980

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022). "Post-selection inference." *Annual Review of Statistics and Its Application*, 9(1): 505–527. MR4394918. doi: https://doi.org/10.1146/annurev-statistics-100421-044639. 979

Lavine, M. and Schervish, M. J. (1999). "Bayes factors: What they are and what they are not." *The American Statistician*, 53(2): 119–122. MR1707756. doi: https://doi.org/10.2307/2685729. 980

Ly, A. and Wagenmakers, E.-J. (2022). "Bayes factors for peri-null hypotheses." *Test*, 31(4): 1121–1142. MR4517127. doi: https://doi.org/10.1007/s11749-022-00819-w. 982

Matthews, R. (2021). "The p-value statement, five years on." *Significance*, 18(2): 16–19. 979

Nester, M. R. (1996). "An applied statistician's creed." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4): 401–410. MR1423738. doi: https://doi.org/10.2307/2986064. 980

Pittelkow, M.-M., Linde, M., de Vries, Y. A., Hemkens, L. G., Schmitt, A. M., Meijer, R. R., and van Ravenzwaaij, D. (2024). "Strength of statistical evidence for the efficacy of cancer drugs: a Bayesian re-analysis of randomized trials supporting FDA approval." *Journal of Clinical Epidemiology*, 111479. 982

Rouder, J. N. and Morey, R. D. (2012). "Default Bayes factors for model selection in regression." *Multivariate Behavioral Research*, 47(6): 877–903. 981

Schwaferts, P. and Augustin, T. (2020). "Bayesian decisions using regions of practical equivalence (ROPE): Foundations." *pre-print at*: https://epub.ub.uni-muenchen.de/. 981

Shao, J. (2008). *Mathematical Statistics*. Springer Science & Business Media. 982

Tendeiro, J. N., Kiers, H. A., Hoekstra, R., Wong, T. K., and Morey, R. D. (2024). "Diagnosing the misuse of the Bayes factor in applied research." *Advances in Methods and Practices in Psychological Science*, 7(1). doi: https://doi.org/10.1177/25152459231213371. 983

Womack, A. J., León-Novelo, L., and Casella, G. (2014). "Inference from intrinsic Bayes' procedures under model selection and uncertainty." *Journal of the American Statistical Association*, 109(507): 1040–1053. MR3265679. doi: https://doi.org/10.1080/01621459.2014.880348. 981

Young, G. A., Smith, R. L., and Smith, R. L. (2005). *Essentials of Statistical Inference*, volume 16. Cambridge University Press. MR2170828. doi: https://doi.org/10.1017/CBO9780511755392. 982