A Cautionary Note on Estimating Effect Size

Don van den Bergh[*][1], Julia M. Haaf[1], Alexander Ly[1,2],

Jeffrey N. Rouder[3], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam

[2]Centrum Wiskunde & Informatica

[3]University of California Irvine

Abstract

An increasingly popular approach to statistical inference is to focus on the estimation of effect size while ignoring the null hypothesis that the effect is absent. We demonstrate how this common "null hypothesis neglect" may result in effect size estimates that are overly optimistic. The overestimation can be avoided by incorporating the plausibility of the null hypothesis into the estimation process through a "spike-and-slab" model.

Consider the following hypothetical scenario: a colleague from the biology department has just conducted an experiment and approaches you for statistical advice. The analysis yields $p < 0.05$ and your colleague believes that this is grounds to reject the null hypothesis. In line with recommendations both old (e.g., Grant, 1962; Loftus, 1996) and new (e.g., Cumming, 2014; Harrington et al., 2019) you convince your colleague that it is better to replace the p-value

with a point estimate of effect size and a 95% confidence interval (but see Morey et al., 2016). You also manage to convince your colleague to plot the data (see Figure 1). Mindful of the reporting guidelines of the *Psychonomic Society*[1] and *Psychological Science*[2], your colleague reports the result as follows: "Cohen's $d = 0.30$, CI $= [0.02, 0.58]$".
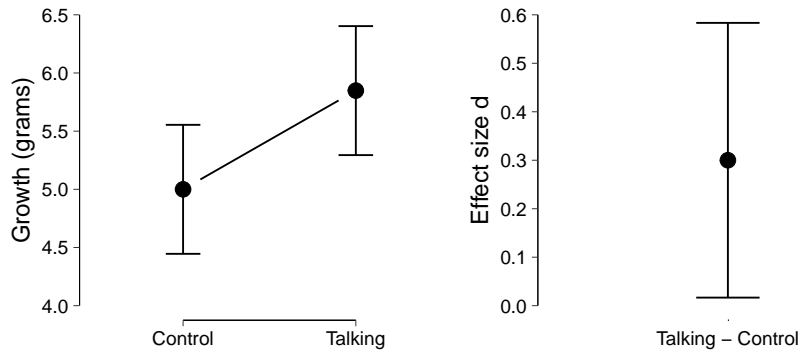

Figure 1: Standard estimation results for the fictitious plant growth example. Left panel: a descriptives plot with the mean and 95% confidence interval of plant growth in the two conditions. Right panel: point estimate and 95% confidence interval for Cohen's *d*.

Based on these results, what would be a reasonable point estimate of effect size? A straightforward and intuitive answer is "0.30". However, your colleague now informs you of the hypothesis that the experiment was designed to assess: "plants grow faster when you talk to them".[3] Suddenly, a population effect size of "0" appears eminently plausible. Any observed difference may merely be due to the inevitable sampling variability.

The example above is, of course, rhetorical. In practice, hypotheses and expectations should be specified up front. Nevertheless, the example raises the question: When are effect sizes overestimated? Standard point estimates and confidence intervals ignore the possibility that the effect is spurious (i.e., the null hypothesis $\mathcal{H}_0$). This is not problematic when $\mathcal{H}_0$ is deeply implausible, either because $\mathcal{H}_0$ was highly unlikely *a priori* or because the data decisively

---

[1] https://www.springer.com/psychology?SGWID=0-10126-6-1390050-0
[2] https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STAT
[3] This example is inspired by Berger and Delampady (1987).

undercut $\mathcal{H}_0$. But when the data fail to undercut $\mathcal{H}_0$, or when $\mathcal{H}_0$ is highly likely *a priori* (i.e., "plants do not grow faster when you talk to them"), then $\mathcal{H}_0$ is not ruled out as a plausible account of the data. Effect size estimates that ignore a plausible $\mathcal{H}_0$ are generally overly optimistic and overconfident: the fact that $\mathcal{H}_0$ provides an acceptable account of the data should shrink effect size estimates towards zero.

In this paper, we discuss the spike-and-slab model and its merits to avoid overestimating effect size. First, we formally introduce the spike-and-slab model. Second, we apply the spike-and-slab model to the example in the introduction and illustrate how it circumvents overestimation. Third, we visualize how the spike-and-slab model may shrink the estimated effect size toward zero in general. Fourth, we demonstrate the spike-and-slab model by reanalyzing the data of Heycke et al. (2018). Finally, we conclude with practical recommendations and a discussion on when to use the spike-and-slab model.

A Spike-and-Slab Perspective

The spike-and-slab approach has been broadly discussed in the statistical literature (Geweke, 1996; Ishwaran, Rao, et al., 2005; O'Hara, Sillanpää, et al., 2009) and psychological literature (Bainter et al., 2020; Yu et al., 2018). Conceptually, the approach is relatively straight-forward.

This section formalizes the conceptual ideas from the introduction in the spike-and-slab model (Clyde et al., 1996; Mitchell & Beauchamp, 1988; Rouder et al., 2018). The population effect size is typically approximated with a sample estimate. This sample estimate assumes that $\mathcal{H}_0$ is false and that the population effect size is nonzero. To formalize this, let δ denote the population effect size, let $\hat{\delta}$ denote an estimate, and let $\hat{\delta} \mid \mathcal{H}_1$ denote an estimate that assumes that $\mathcal{H}_1$ is true. Assuming that $\mathcal{H}_0$ is true leads to $\hat{\delta} \mid \mathcal{H}_0$, which is usually 0. Key is that both estimates, $\hat{\delta} \mid \mathcal{H}_1$ and $\hat{\delta} \mid \mathcal{H}_0$, are *conditional* on the hypotheses. For example, $\hat{\delta} \mid \mathcal{H}_1$ should be read as "the estimated effect size under the

alternative hypothesis".

The first component, the spike, corresponds to the position that talking to plants does not affect their growth (i.e., $\delta = 0$), whereas the second component, the slab, corresponds to the position that speaking to plants does affect their growth (i.e., $\delta \neq 0$). The spike and slab are analogous to $\mathcal{H}_0$ and $\mathcal{H}_1$ discussed above. Both components are deemed *a priori* equally likely, such that the prior probability for each component is 1/2. One can also assign prior probabilities other than 1/2, if this can be motivated by prior research or theories (e.g., Wilson & Wixted, 2018). After seeing the data, the prior probabilities of each component, $\Pr(\text{spike})$ and $\Pr(\text{slab})$, are updated to posterior probabilities, $\Pr(\text{spike} \mid \text{data})$ and $\Pr(\text{slab} \mid \text{data})$.

Quantifying the uncertainty about the two components using probabilities allows us to account for the uncertainty of each component. Furthermore, this allows us to consider the *marginal* estimate of effect size. The marginal estimate of the spike-and-slab model weighs the estimates of each component by its plausibility after seeing the data. That is, we sum the estimates conditional on each component weighed by the posterior probabilities of the components:

$$\hat{\delta} = \left(\hat{\delta} \mid \text{spike} \right) \Pr(\text{spike} \mid \text{data}) + \left(\hat{\delta} \mid \text{slab} \right) \Pr(\text{slab} \mid \text{data}).$$

The result is a *marginal* estimate where the estimate of the slab is shrunken toward that of the spike (or vice versa). To obtain a marginal estimate of effect size, we need to choose an inferential paradigm. In a frequentist approach, one can use penalized maximum likelihood. This foregoes marginalizing over the spike and slab and instead directly adds a penalty term to the likelihood to shrink estimates towards zero. Well-known examples are LASSO and ridge regression (Tibshirani et al., 2005). One frequentist approach that does marginalize uses information criteria to derive posterior probabilities for the spike and slab. For example, take the Akaike Information Criterion (AIC; Akaike, 1973) and define $\Pr(\text{spike} \mid \text{data}) = {}^{\text{AIC|spike}}/_{(\text{AIC|spike}+\text{AIC|slab})}$. This

approach is taken by, for instance, Burnham and Anderson (2002).

An alternative is a Bayesian approach. Here we assign probability distributions to unknown parameters, such as effect size. Even though choosing Bayesian analyses requires the specification of additional prior distributions, the advantage is that the posterior model probabilities follow directly from Bayes theorem. Whereas frequentist approaches need to specify a penalty term or choose an information criterion, Bayesian approaches need to specify prior distributions on the model parameters (e.g., δ) and the models themselves (Pr (spike) and Pr (slab)).

However, a Bayesian approach provides a posterior distribution for effect size which fully captures the uncertainty over possible effect sizes and allows for easy computation of uncertainty intervals. In the remainder of this paper, we adopt the Bayesian approach for the spike-and-slab model. For a full derivation of the posterior distribution, see Appendix A.

To illustrate both the overestimation and the spike-and-slab model we re-analyze the fictitious data from Figure 1. R code for the analysis is available at https://osf.io/uq8st/. In almost all current empirical work, an estimate of effect size is based solely on the slab, which yields a point estimate and an uncertainty interval (for frequentists, $\hat{\delta} = 0.30$, 95% CI: [0.02, 0.58]; for Bayesians $\hat{\delta} = 0.29$, 95% CRI: [0.02, 0.57]). The spike-and-slab model, however, also considers the possibility that an effect can be absent; consequently, the overall estimate from the spike-and-slab model is a weighted average of the two components, shrinking the estimate towards zero. Figure 2 contrasts the slab-only estimation against the spike-and-slab estimation.

Compared to the traditional results based only on the slab, the posterior mean and central 95% credible interval of the spike-and-slab model are shrunken towards 0 (i.e., 0.14 (95% CRI: [0.00, 0.48]) vs. 0.29 (95% CRI: [0.02, 0.57])). This shrinkage is due to the non-negligible probability that the effect is absent. The spike-and-slab posterior represents the plausibility that the effect is absent by the height of the spike, and the uncertainty about the effect's magnitude,
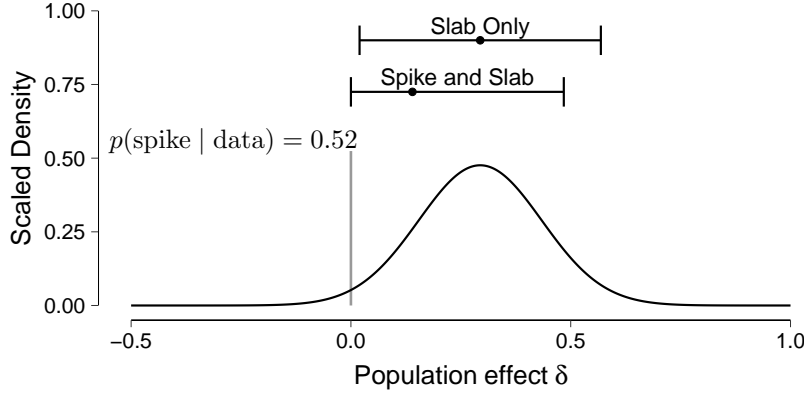
Figure 2: The spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\hat{\delta} = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\text{slab} \mid \text{data}) = 0.48$). The grey line represents the posterior probability of the spike (i.e., $\hat{\delta} = 0$: the effect is absent). The error bars and dots above the density show 95% credible intervals and the posterior mean for the slab-only model and for the spike-and-slab model.

given that it is present, by the width of the slab. Note that as the posterior probability of the spike decreases, the spike-and-slab results approach those of the slab-only model.

The Influence of the Spike

In the fictitious example, the spike-and-slab model avoids overestimation by shrinking estimates of effect size towards zero. That result may not be surprising, as the effect was small. However, it makes one wonder to what extent the spike-and-slab model helps with estimation. What differs between a slab only model and the spike-and slab? In this section, we illustrate how the estimated effect size shrinks towards zero in general. We visualize the shrinkage as a function of the observed effect size, the prior on the standard deviation of effect size under the slab, the sample size, and the prior probability of the spike. We chose these parameters because the posterior distribution is fully determined by these quantities (see Appendix A).

Figure 3 shows the relation between the observed effect size and the esti-

6

mated effect size for the slab and for the spike-and-slab for 40 observations and 100 observations. All plots show that a smaller prior standard deviation of the slab induces more shrinkage towards zero. This makes sense as a small prior standard deviation implies there is more prior mass near the mean of the prior, which is zero. The influence of the prior standard deviation is intrinsic to a Bayesian approach, but not to the spike-and-slab model. Comparing the plots between the two columns illustrates the influence of the spike; whenever the observed effect size is near zero, the estimate is shrunken towards zero in the right column but not in the left column. However, when the observed effect size is far from zero, there is little shrinkage. The shrinkage can be explained in the following way. Whenever the observed effect size is small, the data are well described by an effect size of zero and thus the posterior probability of the spike is substantial. As a result the marginal estimate is shrunken towards the spike's estimate, 0. In contrast, when the observed effect size is large the data are poorly described by an effect size of zero and the posterior probability of the spike is negligible. As a consequence, the estimate of the spike-and-slab is practically equivalent to the estimate of the slab. The plots in the right column of Figure 3 shows the effect of sample size on the shrinkage. For the bottom right plot, $N = 100$, if the observed effect size is small then the estimate is still shrunken towards 0, but as the observed effect size grows the shrinkage decreases much more quickly than in the top right plot where $N = 40$. This makes sense from a signal-detection perspective. If the observed effect size is, for example, 0.3 after 40 observations, the posterior probability of the spike is substantial. However, after collecting 60 additional observations while the observed effect size remains 0.3, the posterior probability of the spike decreases as it becomes increasingly less probable that the data generating model had an effect size of zero.

Next, we explore the relationship between shrinkage and the prior probability of the spike. Figure 4 shows the shrinkage for various prior probabilities. The smaller the prior probability of the spike, the less the effect size is shrunken
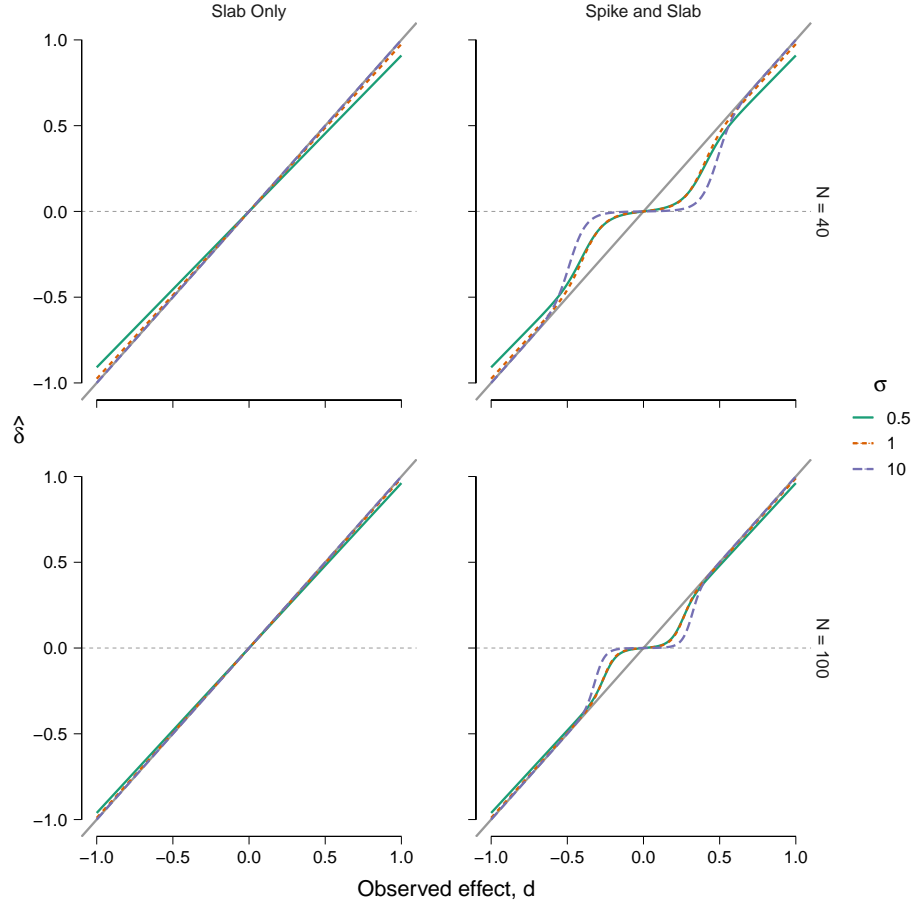
Figure 3: Observed effect size versus posterior mean for different model components and prior standard deviations. The left column shows inference based on the slab only model while the right column shows inference based on the spike-and-slab model. In the top row, the sample size was 40 while in the bottom row the sample size was 100. Different lines represent different standard deviations for the prior distribution on δ. The prior probability of the spike was $1/2$. Inspired by Figure 5 of Rouder et al. (2018).

towards 0. If the prior probability is small then the spike was a-priori implausible and less evidence is needed to make its influence negligible.

Empirical Example: Reanalysis of Two Minds

To improve upon the previous fictional example, we highlight how the spike-and-slab approach can be used in psychological practice by reanalyzing the results of
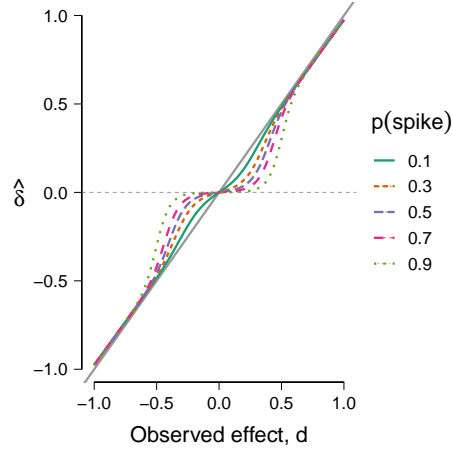
Figure 4: Observed effect versus posterior mean (x-axis) versus the posterior mean of the spike-and-slab model (y-axis). The different lines represent different standard prior probabilities of the spike. The figure is based on 40 observations with a prior standard deviation of 1.

Heycke et al. (2018). Heycke et al. (2018) conducted two registered replications of Rydell et al. (2006). We first briefly explain the design of the study before reanalyzing the results, for a detailed description see the "Procedure" section in Heycke et al. (2018). Afterward, we reanalyze the Explicit Evaluation and Implicit Evaluation analyses with a spike-and-slab model. Finally, we provide a robustness analysis.

Participants were shown a positive or negative prime on a computer screen followed by an image of a person. Next, several behavioral descriptions that were either negative or positive appeared underneath the image of the person. Participants were asked to indicate whether the behavioral description was characteristic or uncharacteristic for the person shown. After the first block of trials, participants explicitly evaluated the target person and performed an implicit association task (IAT). In total, data of 51 participants was analyzed. Heycke et al. (2018) found that an explicit description had an effect while the IAT showed an indication of an effect.

Explicit Evaluation   In the analysis of the explicit evaluations, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that the rating of the

target character is more positive if positive information is shown before negative information: $t(27) = 11.52$, $p < .001$; $\text{BF}_{10} = 1.37 \times 10^9$, $d = 2.09$, 95% HDI $[1.41, 2.79]$.[4] The magnitude of the effect is large and thus a spike-and-slab reanalysis yields practically the same results: $\hat{\delta} = 2.10$, 95% CRI: $[1.74, 2.47]$.[5]

Implicit Evaluation   In the analysis of the IAT, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that when negative primes were presented before positive primes there was some indication that the IAT rating became more negative: $t(27) = -2.54$, $p = .017$, $\text{BF}_{10} = 2.92$, $d = -0.44$, 95% HDI $[-0.83, -0.06]$. Here, the magnitude of the effect is smaller and as a consequence the results from the spike-and-slab reanalysis are more conservative: $\hat{\delta} = -0.35$, 95% CRI: $[-0.75, 0.00]$. The estimate of effect size is shrunken towards 0 because the spike provides a reasonable account of the data, $\Pr(\text{spike} \mid \text{data}) = 0.25$.

Robustness analysis   In the reanalyses above the prior probability of the spike was set to 0.5. One might wonder how robust or how volatile the results are to changes in the prior probability of the spike. Figure 5 visualizes the influence of the prior on the spike. In the left panel that shows the explicit evaluation data, the different estimates for different prior probabilities are practically identical. For this analysis, the data dominate the prior. In contrast, in the right panel that shows the explicit evaluation data, the prior probability of the spike has a large impact on the results. Here, the data are less informative and the prior has more influence.

Don: call this adaptive shrinkage

---

[4]These are the statistics reported by Heycke et al. (2018). BF stands for Bayes factor, see also appendix A. HDI is short for highest density interval, a type of credible interval.

[5]The difference between the point estimate and the credible intervals is possibly caused by the difference in prior distributions for effect size. Heycke et al. (2018) use a Cauchy prior whereas we use a normal prior.
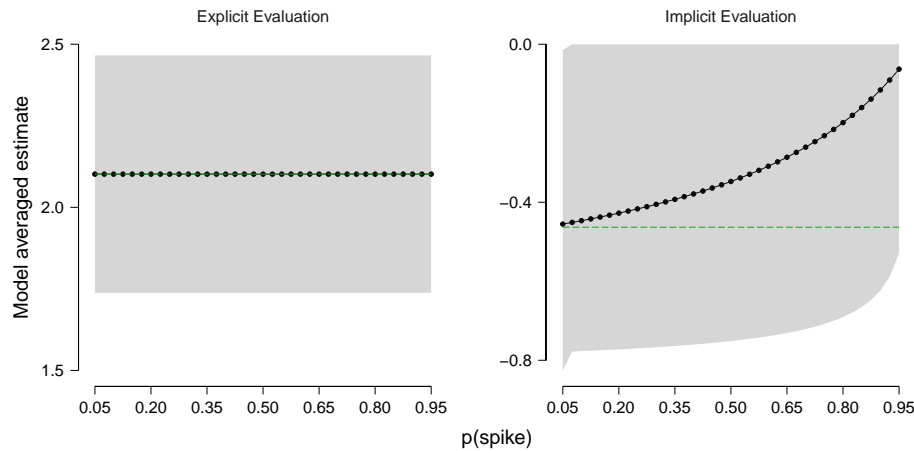
Figure 5: Robustness analysis that shows the prior probability of the spike (x-axis) versus spike-and-slab estimates (y-axis) for the explicit evaluation (left panel) and the implicit evaluation (right panel). Solid points show the point estimate of the spike-and-slab and the gray area represents the accompanying 95% credible interval. The green horizontal dashed line shows the estimate of the slab.

Discussion

Standard estimates of effect size ignore the null hypothesis and are therefore overconfident, that is, farther from from zero than they should be. The spike-and-slab model remedies this problem by explicitly considering the possibility that an effect is absent (Robinson, 2019; Rouder et al., 2018). The core idea dates back to Jeffreys (1939; see also Ly and Wagenmakers, 2020); nonetheless, it is largely ignored in empirical practice, in statistical education, and in journal guidelines. We believe the spike-and-slab model is a useful addition to existing statistical tools of researchers and that it will make the interpretation of the estimates, even those from a preregistered study, more robust across different models.

What if All Null Hypotheses Are False?

The spike-and-slab approach clashes with the popular estimation mindset, where it is argued that statistical significance should be abandoned in favor of estimation (Cumming, 2014; Cumming & Calin-Jageman, 2016; McShane et al.,

2019; Valentine et al., 2015). One argument to forgo hypothesis testing is that all null hypotheses are false (Cohen, 1990; Meehl, 1978) and therefore there is no need to consider a component that states that an effect is exactly zero. The statistical counterargument is that, even if point null hypotheses are false, they are still mathematically convenient approximations to more complex hypotheses that allow mass on an interval close to zero (i.e., perinull hypotheses; Berger & Delampady, 1987; George & McCulloch, 1993; Ly et al., 2020). Thus, from a pragmatic perspective it is irrelevant whether or not null hypotheses are exactly true: in the spike-and-slab model, a narrow interval around zero will shrink estimates towards zero almost as much as the point null spike component will.

When to Ignore the Spike

There are two scenarios in which the presence of the spike can safely be ignored. First, the spike may be deeply implausible. This happens most often in problems of pure estimation, such as when determining the relative popularity of two politicians or the proportion of Japanese cars on the streets of New York. In such cases, no value or interval needs to be singled out for special attention. Second, the data, or data from prior studies, may provide overwhelming evidence that an effect is present, as in the reanalysis of the Explicit Evaluation data. When this happens, the results from a spike-and-slab model become virtually identical to those of a slab-only model, and the inclusion of the spike does not offer an added benefit albeit that the spike also does not hurt.

Conclusion

Standard methods for estimating effect size produce results that are overly optimistic. This tendency toward high estimates can be corrected by applying the spike-and-slab model which explicitly accounts for the possibility that the effect is absent. The spike-and-slab approach is not meant as a tool to downplay other researchers' findings that one does not believe in. Instead, it provides a more robust estimate of the size of an effect of high-quality studies if null

and alternative hypothesis are plausible. We believe that approach allows researchers a more nuanced interpretation of their own results taking into account the plausibility that there is any effect.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science*, *3*(1), 66–80.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.)* Springer Verlag.

Clyde, M. A., Desimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, *91*(435), 1197–1208.

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*, 80–101.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304–1312.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith)*, 609–620.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.-L. T., Drazen, J. M., & Hamel, M. B. (2019). New guidelines for statistical reporting in the journal.

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? a registered replication of rydell et al.(2006). *Cognition and Emotion*, *32*(8), 1708–1727.

Ishwaran, H., Rao, J. S. et al. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharskỳ, Š., Derks, K., Gronau, Q. F., Komarlu Narendra Gupta, A. R., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p-value hypothesis test. *Computational Brain & Behavior*, (3), 153–161.

Ly, A., & Wagenmakers, E.-.-J. (2020). Bayes factors for peri-null hypotheses. *Manuscript in preparation*.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Atatistician*, *73*(sup1), 235–245.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

O'Hara, R. B., Sillanpää, M. J. et al. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian analysis*, *4*(1), 85–117.

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197.

Yu, C.-H., Prado, R., Ombao, H., & Rowe, D. (2018). A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fmri. *Journal of the American Statistical Association*, *113*(524), 1395–1410.

A: Derivation of the Posterior Distribution

Here we derive the posterior distribution for the spike-and-slab model. We first derive the results for the slab and spike individually and afterward we combine them. Using the slab, we explain the model and assign a prior distribution to effect size. Afterward, we use Bayes' theorem to obtain the posterior distribution for the parameters.

In the example, we use a paired samples t-test. Following Rouder et al. (2018), we assume that the differences between the paired samples, denoted Z_i, are normally distributed with unknown mean δ and a known variance of 1. As prior distribution for δ we use a normal distribution with mean 0 and unknown variance σ^2. This implies $Z_i \sim \mathcal{N}(\delta, 1)$ for the data and $\delta \sim \mathcal{N}(0, \sigma^2)$ for the prior. The posterior distribution for δ is obtained through Bayes' theorem:

$$\underbrace{p(\delta \mid \boldsymbol{Z})}_{\substack{\text{Posterior} \\ \text{distribution}}} = \underbrace{p(\delta)}_{\substack{\text{Prior} \\ \text{distribution}}} \times \frac{\overbrace{L(\boldsymbol{Z} \mid \delta)}^{\text{Likelihood}}}{\underbrace{p(\boldsymbol{Z})}_{\substack{\text{Marginal} \\ \text{Likelihood}}}}.$$

For the prior distribution we use a normal distribution with mean 0 and variance σ^2. The likelihood is given by:

$$L(\boldsymbol{Z} \mid \delta, \text{slab}) = \prod_{i=1}^{N} \mathcal{N}(Z_i \mid \delta, 1),$$

$$= (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{Z} + s_{\boldsymbol{Z}}^2 + \delta^2 - 2\bar{Z}\delta\right)\right),$$

where $\bar{\boldsymbol{Z}}$ and $s_{\boldsymbol{Z}}^2$ are the sample mean and sample variance of Z_i respectively. Next, we compute the marginal likelihood by integrating out the likelihood times prior with respect to δ.

$$p\left(\boldsymbol{Z} \mid \text{slab}\right) = \int_{-\infty}^{\infty} L(\boldsymbol{Z} \mid \delta) p\left(\boldsymbol{Z}\right) \, \mathrm{d}\delta,$$

$$= (2\pi)^{-\frac{N+1}{2}} \exp\left(-\frac{N}{2}\left(\bar{Z} + s_{\boldsymbol{Z}}^2\right)\right),$$

$$\times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{Z}}{\sigma^2}\right)\right) \, \mathrm{d}\delta.$$

Here we may recognize a Gaussian integral and use the following identity:

$$\int_{-\infty}^{\infty} \exp\left(-ax^2 + bx + c\right) \, \mathrm{d}x = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c\right).$$

Filling in the identity and simplifying yields:

$$p\left(\boldsymbol{Z} \mid \text{slab}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{Z} + s_{\boldsymbol{Z}}^2\right)\right) \frac{\exp\left(\frac{N^2\bar{Z}^2}{2\left(N + \frac{1}{\sigma^2}\right)}\right)}{\sqrt{N + \frac{1}{\sigma^2}}}.$$

Next, we can obtain an expression for the posterior distribution. However, often it suffices to write out the expression for the likelihood times prior and then (somehow) recognize a known distribution. This is particularly common in Gibbs sampling where one is interested in the conditional posterior distributions. We also do this here, as it reduces inference about the posterior distribution (e.g., what is the mean or variance) to inference about a known distribution, in this case a normal distribution:

$$p\left(\delta \mid \boldsymbol{Z}, \text{slab}\right) \propto (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{Z} + s_{\boldsymbol{Z}}^2\right)\right) \exp\left(-\frac{N}{2}\left(\delta^2 - 2\bar{Z}\delta\right)\right),$$

$$\times (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\delta^2\right),$$

$$\propto \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{Z}}{\sigma^2}\right)\right).$$

We recognize a normal distribution with variance $\sigma_1^2 = \frac{1}{N + \frac{1}{\sigma^2}}$ and mean $\mu_1 = N\bar{Z}\sigma_1^2$. Thus we have $p\left(\delta \mid \boldsymbol{Z}\right) \propto \mathcal{N}\left(\mu_1, \sigma_1^2\right)$.

Next we compute the same for the spike. The spike states that $Z_i \sim \mathcal{N}\left(0, 1\right)$

and contains no parameters to estimate. Thus there are no prior distributions to specify and the only thing we need to compute is the marginal likelihood:

$$p\left(\boldsymbol{Z} \mid \text{spike}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s_{\boldsymbol{Z}}^2\right)\right).$$

Using both marginal likelihoods we can now obtain the Bayes factor in favor of the spike:

$$\text{BF}_{01} = \frac{p\left(\boldsymbol{Z} \mid \text{spike}\right)}{p\left(\boldsymbol{Z} \mid \text{slab}\right)} = \frac{\sqrt{N + \frac{1}{\sigma^2}}}{\exp\left(\frac{N^2 \bar{\boldsymbol{Z}}^2}{2\left(N + \frac{1}{\sigma^2}\right)}\right)}.$$

We may now obtain inference for both the spike and the slab by computing the posterior probability of the slab:

$$\Pr\left(\text{slab} \mid \boldsymbol{Z}\right) = \frac{\Pr\left(\text{spike}\right)}{\Pr\left(\text{spike}\right) + (1 - \Pr\left(\text{spike}\right))\text{BF}_{01}}.$$

It follows that the cumulative distribution function for the spike-and-slab posterior is given by:

$$P(\delta \leq x \mid \boldsymbol{Z}) = \begin{cases} \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \Phi(x; \mu_1, \sigma_1) & \text{if } x < 0, \\ \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \Phi(x; \mu_1, \sigma_1) & \text{if } x \geq 0, \end{cases}$$

where $\Phi(x; \mu_1, \sigma_1)$ is the cumulative normal distribution. Due to the discontinuity at $x = 0$ there is no useful closed form expression for the posterior density. Nevertheless, the posterior mean of the spike-and-slab model is available in closed form. Using the law of total probability, we have:

$$p(\delta \mid \boldsymbol{Z}) = \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) p(\delta \mid \text{spike}, \boldsymbol{Z}) + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) p(\delta \mid \text{slab}, \boldsymbol{Z}).$$

Computing the mean of left hand side yields:

$$\int_{-\infty}^{\infty} \delta\, p(\delta \mid \boldsymbol{Z})\ \mathrm{d}\delta = \Pr\left(\text{spike} \mid \boldsymbol{Z}\right) \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{spike}, \boldsymbol{Z})\ \mathrm{d}\delta,$$

$$+ \Pr\left(\text{slab} \mid \boldsymbol{Z}\right) \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{slab}, \boldsymbol{Z})\ \mathrm{d}\delta,$$

$$= 0 + \Pr\left(\text{slab} \mid \boldsymbol{Z}\right)\left(\mu_\delta \mid \text{slab}, \boldsymbol{Z}\right).$$

Here $\left(\mu_\delta \mid \text{slab}, \boldsymbol{Z}\right)$ is the posterior mean of effect size under the slab. In a similar fashion, other statistics may be obtained. However, it is also possible to draw samples from marginal posterior distribution. To obtain a sample s, first draw u from a uniform distribution on $[0, 1]$. If $u < \Pr\left(\text{slab} \mid \boldsymbol{Z}\right)$ draw s from $p(\delta \mid \text{slab}, \boldsymbol{Z})$, otherwise s is zero. This approach often used when the integrals become too unwieldy to compute analytically. For example, the R package BAS uses this procedure to compute credible intervals (Clyde et al., 2011).