

Running head: PARAMETER ESTIMATION AND BAYES FACTORS

Bayesian Inference for Psychology, Part IV: Parameter Estimation and Bayes factors.

Jeffrey N. Rouder

University of Missouri

Joachim Vandekerckhove

University of California, Irvine

Jeff Rouder

rouderj@missouri.edu

Abstract

In the psychological literature, there are two seemingly different approaches to inference: that from estimation of posterior intervals and that from Bayes factors. We provide an overview of each method and show that a salient difference is the choice of models. The two approaches as commonly practiced can be unified with a certain model specification, now popular in the statistics literature, called *spike-and-slab* priors. A spike-and-slab prior is mixture of a null model, the spike, with an effects model, the slab. The estimate of the effect size here is a function of the Bayes factor showing that estimation and model comparison can be unified. The salient difference is that common Bayes factor approaches provides for privileged consideration of theoretically useful parameter values, namely, the zero value corresponding to the null, while common estimation approaches do not. Both approaches, either privileging the null or not, are useful depending on the goals of the analyst.

Bayesian Inference for Psychology, Part IV: Parameter Estimation and
Bayes factors.

Bayesian analysis has become increasingly popular in many fields including psychological science. There are many advantages to the Bayesian approach. Some champion its clear philosophical underpinnings where probability is treated as a statement of belief and the focus is on updating beliefs rationally in face of new data (de Finetti, 1974). Others note the practical advantages—Bayesian analysis often provides a tractable means of solving difficult problems that remain intractable in more conventional frameworks (Gelman, Carlin, Stern, & Rubin, 2004). This advantage is especially pronounced in psychological science where substantive models are designed to account for mental representation and processing. As a consequence, the models tend to be complex and nonlinear, and may include multiple sources of variation (Kruschke, 2011b; Lee & Wagenmakers, 2013; Rouder & Lu, 2005). Bayesian analysis, particularly Bayesian nonlinear hierarchical modeling, has been particularly successful at providing straightforward analyses in these otherwise difficult settings (e.g., Vandekerckhove, Tuerlinckx, & Lee, 2011).

Bayesian analysis is not a unified field, and Bayesian analysts disagree with one another in important ways.¹ We highlight here two popular Bayesian approaches that seem incompatible and discuss them in the context of the simple problem of determining whether performance in two experimental conditions differs. In one approach, termed here the **estimation approach**, the difference between the conditions is represented by a parameter, and posterior beliefs about this parameter are updated using Bayes' Rule. From these posteriors, researchers may observe directly which parameter values are plausible, and, as importantly, which are implausible. Two examples are provided in Figure 1. In Figure 1A the posterior distribution is compact in extent and localized well away from zero, and this localization serves as evidence for a substantial

difference between the two conditions. In Figure 1B, in contrast, the value of zero is well within the belly of posterior, indicating that there is little evidence for such a difference. Perhaps the leading advocate of the Bayesian estimation school in psychology is Kruschke (Kruschke, 2011a, 2012). Although the posterior-estimation approach seems straightforward, it is not recommended by a number of Bayesian psychologists including Dienes (2014), Gallistel (2009), Rouder, Speckman, Sun, Morey, & Iverson (2009), and Wagenmakers (2007). These authors instead advocate a **Bayes factor** approach. In Bayesian analysis, it is possible to place probability on models themselves without recourse to parameter estimation. In this case, a researcher could construct two models: one that embeds no difference between the conditions and one that embeds some possible difference. The researcher starts with prior beliefs about the models and then updates these rationally with Bayes' rule to yield posterior beliefs. Evidence from data is how beliefs about the models themselves change in light of data; there may be a favorable revision for either the effects or null-effects model.

Estimation and Bayes factor approaches do not necessarily lead to the same conclusions. Consider for example the posterior in Figure 1A where the posterior credible interval does not include zero. This posterior seemingly provides positive evidence for an effect. Yet, the Bayes factor, which is discussed at length subsequently, is 2.8-to-1. If we had started with 50-50 beliefs about an effect (vs. a lack of an effect), we end up with just less than 75-25 beliefs in light of data. While this is some revision of belief, this small degree is considered rather modest (Jeffreys, 1961; Raftery, 1995).

This divergence leaves the nonspecialist in a quandary about whether to use estimation or Bayes factors. Faced with this quandary, we fear that some will ignore Bayesian analysis altogether. In this paper we address this quandary head on: First we first draw a sharp contrast between the two approaches and show that they provide for quite different views of evidence. Then, to help understand these differences, we provide a

unification. We show that the Bayes factor may be represented as estimation under a certain model specification known in the statistics literature as a *spike-and-slab* model (George & McCulloch, 1993). With this demonstration, one difference between estimation and a Bayes factor approach comes into full view—it is a difference in model specification rather than any deep difference in the Bayesian machinery. These spike-and-slab models entail different commitments than more conventional models. Once researchers understand these different commitments, they can make informed and thoughtful choices about which are most appropriate for specific research applications.

Estimation

Bayesian estimation is performed straightforwardly through updating by Bayes' rule. Let us take a simple example where a set of participants provide performance scores in each of two conditions. For example, consider a priming task where the critical variable is the response time, and participants provide a mean response time in a primed and unprimed condition. Each participant's data may be expressed as a difference score, namely the difference between mean response times. Let Y_i , $i = 1, \dots, n$ be these difference scores for n participants. In the usual analysis, researchers would perform a t -test to assess whether these difference scores are significantly different than zero.

Bayesian analysis begins with consideration of a model, and in this case, we assume that each difference score is a draw from a normal with mean μ and variance σ^2 :

$$Y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

In the following development, we will assume that σ^2 is known to simplify the exposition, but it is straightforward to dispense with this assumption. It is helpful to consider the model in terms of effect sizes, δ , where $\delta = \mu/\sigma$ is the true effect size and is the parameter

of interest.

Bayesian analysis proceeds by specifying beliefs about the effect-size parameter δ . The beliefs are expressed as a prior distribution on parameters. In this article, we use the term *prior* and *model* interchangeable as a prior is nothing more than a model on parameters. Model \mathcal{M}_1 provides prior beliefs on δ .

$$\mathcal{M}_1 : \quad \delta \sim \text{Normal}(0, \sigma_0^2). \quad (2)$$

The centering of the distribution at zero is interpreted as a statement of prior equivalence about the direction of any possible effect—negative and positive effects are *a priori* equally likely. The prior variance, σ_0^2 must be set before analysis, and it is helpful to explore how the value of this setting affects estimation. Figure 2A shows this effect. Ten hypothetical values of Y_i , the difference scores, are shown as blue segments across the bottom of the plot. The sample mean of these ten is shown as the vertical line. The posterior beliefs about δ are shown for three different prior settings. The first prior setting, $\sigma_0 = .5$, codes an *a priori* belief that δ is not much different than zero. The second prior setting, $\sigma_0 = 2$, is a fairly wide setting that allows for a large range of reasonable effect sizes without mass on exceedingly large values. The third prior setting, $\sigma_0 = 5000$ indicates that researcher is unsure of the effect size, and holds the possibility that it can be exceedingly large. Even though the priors are quite different, the posteriors distributions, shown, are quite similar. We may say that the posterior is robust to wide variation in prior settings. In fact, it is possible to set $\sigma_0 = \infty$ to equally weight all effect sizes *a priori*, that is to not make any prior commitment at all, and in this case, the posterior would be indistinguishable from that for $\sigma_0 = 5000$. This robustness to prior settings is an advantage of Model \mathcal{M}_1 on δ . As will be shown, while this robustness holds for \mathcal{M}_1 , it is not a general property of Bayesian estimation. We will introduce models subsequently where it does not hold.

There are many ways to use the posterior distributions to state conclusions. One could simply inspect them and interpret them as needed (Gelman & Shalizi, 2013 and Rouder et al., 2008 take variants of this approach). Alternatively, one could make a set of inferential rules. In his early career, Lindley (1965) recommended inference by *highest-density credible intervals* (HDCIs). These highest-density credible intervals contain a fixed proportion of the mass, say 95%, and posterior values inside the interval are greater than those outside the interval. Example of these HDCIs are shown in Figure 1 with the dashed vertical lines. Values outside the intervals may be considered sufficiently implausible to be untenable. By this reasoning, there is evidence for an effect in Figure 1A as zero is outside the 95% credible interval. Figure 2B shows that inference by credible intervals does not depend heavily on the prior setting σ_0^2 . Shown is the minimal effect size needed such that zero is excluded from the lower end of the credible interval. As can be seen, this value stabilizes quickly and varies little.

Kruschke (2012) takes a similar approach. A posterior interval may be compared to a pre-established region, called a *region of practical equivalence* or ROPE. ROPES are small intervals around zero that are considered to be practically the same as zero. An example of a ROPE might be the interval on effect sizes from $-.2$ to $.2$. If the HDCI falls completely outside of the ROPE, one concludes that the null hypothesis is false. If the HDCI falls completely inside of the ROPE, one concludes that the null hypothesis is (for all practical purposes) true. Inferences drawn this way are robust to the prior setting of σ_0 , and arbitrarily large (even infinite) values may be chosen.

Bayes Factors

In Bayesian analysis, it is possible to place beliefs directly onto models themselves and update these beliefs with Bayes' rule. Let \mathcal{M}_A and \mathcal{M}_B denote any two models. Let $Pr(\mathcal{M}_A)$ and $Pr(\mathcal{M}_B)$ be *a priori* beliefs about the plausibility of these two models. It is

more desirable to state relative beliefs about the two models as odds. The ratio $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ is the relative plausibility of the models, and for example, the statement $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B) = 3$ indicates that Model \mathcal{M}_A is three times as plausible as Model \mathcal{M}_B . Odds such as $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ are called *prior odds* because they are stipulated before seeing data. They may be contrasted to *posterior odds*, which are the same odds in light of the data and denoted $Pr(\mathcal{M}_A | \mathbf{Y})/Pr(\mathcal{M}_B | \mathbf{Y})$. be the prior and posterior odds, respectively. Bayes rule for updating to posterior odds from prior odds is

$$\frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)} = \frac{f(\mathbf{Y} | \mathcal{M}_A)}{f(\mathbf{Y} | \mathcal{M}_B)} \times \frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)}.$$

The updating factor, $f(\mathbf{Y} | \mathcal{M}_A)/f(\mathbf{Y} | \mathcal{M}_B)$, is called the *Bayes factor*, and it describes how the data have led to a revision of beliefs about the models. Several authors including Jeffreys (1961) and Morey, Romeijn, & Rouder (2016) refer to the Bayes factors as the *strength of evidence from data about the models* precisely because the strength of evidence should refer to how data lead to revision of beliefs. The Bayes factor has a second meaning stemming from it being the probability of data under models. The probability of data may be thought of as the predictive accuracy of a model. The data in this case is the data we obtain in an experiment, and if the probability of data is high, then the model predicted the observed data to be where they were observed. If the probability if data is low, then the model predicted them to be elsewhere. The Bayes factor is the relative predictive accuracy of one model relative to another. The deep meaning of Bayes rule is that the strength of evidence is the relative predictive accuracy, and this equality is captured by the Bayes factor.

We denote the Bayes factor by B_{AB} , where the subscripts indicate which two models are being compared. A Bayes factor of $B_{AB} = 10$ means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_A ; likewise, a Bayes factor of $B_{AB} = .1$

means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_B . Bayes factors of $B_{AB} = \infty$ and $B_{AB} = 0$ correspond to infinite support of one model over the other with the former indicating infinite support for model \mathcal{M}_A and the latter indicating infinite support for model \mathcal{M}_B .

For the simple example of comparing performance in two experimental conditions, we instantiate a separate model for an effect and for an invariance. The previous model, \mathcal{M}_1 given in (2) serves as a good model for an effect. Needed is a null-effect model, and this is given by

$$\mathcal{M}_0 : \quad \delta = 0.$$

With this setup, the Bayes factor is straightforward to compute.²

The Bayes factor is more dependent on the prior setting σ_0^2 than is the posterior distribution under Model \mathcal{M}_1 . Figure 3A shows the effects of increasing σ_0 . As can be seen, the Bayes factor B_{10} favors the alternative when σ_0 is small (say, near 1) but decreases toward zero as σ_0 becomes increasingly large. Arbitrarily diffuse priors on effect size in the alternative leads to arbitrarily strong support for the null model over the alternative (Lindley, 1957), and this result contrasts to that for inference with credible intervals where arbitrarily diffuse priors could be used without any cost. This result occurs because the Bayes factor is sensitive to the complexity of the model, and when the $\sigma_0^2 = \infty$, the alternative can account for all data without constraint. Consequently, it is penalized completely. Figure 3B provides an different view of the effect of prior setting σ_0 . It shows the minimum positive effect size need to support a Bayes factor of 3-to-1 in favor of Model \mathcal{M}_1 over \mathcal{M}_0 and is comparable to Figure 2B. As can be seen, inference by Bayes factor is more sensitive to prior settings than inference by estimation.

At first glance, this dependence of the Bayes factors on the prior settings may seem

undesirable. Researchers can seemingly obtain different results by adjusting the prior settings perhaps undermining the integrity of their conclusions. This dependence seems all the more undesirable when contrasted to the the relative independence of posterior intervals on prior settings as shown in Figure 2. However, the situation is far more nuanced, and we believe researchers should not worry too much about prior dependence or lack thereof. Indeed both estimation and Bayes factor are derivative of the same Bayes rule, and the differences are more subtle and perhaps even more interesting than they first appear. In the next section we provide a unification, and with this unification can pinpoint the differences and make recommendations for researchers.

Unification

The differences between the estimation and Bayes factor approach can be understood by combining Models \mathcal{M}_0 and \mathcal{M}_1 . Figure 4A shows the combination, which is expressed as a mixture. One component of the mixture is the usual normal model on effect size (Model \mathcal{M}_1), and this component is denoted by the curve in Figure 4A. The other component is a placing mass on the point of zero, and this component is denoted by the arrow. In this case, the arrow is half-way up its scale, shown in dashed line, indicating that half of the total mass is placed at zero, and the other half is distributed around zero. This model is well known in the statistics literature as a *spike-and-slab model* (Mitchell & Beauchamp, 1988). We denote it by Model \mathcal{M}_s .³ The spike-and-slab model in has two parameters: the amount of probability in the spike, denoted ρ_0 , and the variance of the slab, denoted σ_0^2 . Figure 4A shows the case where $\rho_0 = 1/2$ and $\sigma_0^2 = 1$.

It is straightforward to update beliefs about δ in the spike-and-slab model using Bayes rule.⁴ Figure 4B-C show a few examples for different observed effect sizes. In all cases, the resulting posterior is in the spike-and-slab form, but the spike has changed mass and the slab has shifted. Figure 4B shows the posterior for a small observed effect size of

0.1. The spike is enhanced as the effect is compatible with a null effect. The slab is attenuated in mass, narrowed, and shifted from 0 to about .1. Figure 4B shows the posterior for a large observed effect size of 0.5. The spike is attenuated as the effect is no longer compatible with the null, and the slab is enhanced, narrowed, and shifted from 0 to about .5.

There is a tight relationship between the spike-and-slab posterior distribution and the Bayes factor B_{01} for the comparison between Model \mathcal{M}_0 and \mathcal{M}_1 . The Bayes factor describes the change in the spike. The prior probability of the spike, ρ_0 , can be expressed as odds, $\omega_0 = \rho_0/(1 - \rho_0)$. The posterior probability of the spike, ρ_1 , can likewise be expressed as odds. The Bayes factor is the change in odds: ω_1/ω_0 . In Figure 4B, for example, the initial odds on the spike were 1-to-1, indicating that equal mass was in the spike as was in the slab. In light of data, the posterior odds were 7.4-to-1, or that 88% of the posterior mass was in the spike and 12% of posterior mass was in the slab. Indeed, the Bayes factor for this case is $B_{01} = 7.4$, and this factor describes the change in odds in the spike in light of data.

The spike-and-slab Model \mathcal{M}_s yields posterior estimates of effect size that behave differently, in fact more advantageously, than the slab-only estimates from \mathcal{M}_1 . Figure 5A-B shows the comparison. The solid curves are posterior means of δ as a function of observed effect size d . For the slab-only specification (Panel A), the estimated mean follows the observed value, and do so for all prior values of σ_0^2 . But, for the spike-and-slab specification (\mathcal{M}_s , Panel B), there is shrinkage toward zero. Shrinkage is well known in hierarchical models and often results in better calibrated estimates (James & Stein, 1961; Efron & Morris, 1977). The shrinkage from the spike-and-slab model is *adaptive* in that shrinkage toward zero is sizable for small observed values while there is hardly any shrinkage for large values. Adaptive shrinkage is an exceedingly useful part of modern Bayesian analysis. It is a Bayesian approach to classification and smoothing, and,

as will be discussed, has become popular in multivariate settings. The amount of adaptive shrinkage is dependent in a reasonable way on the prior setting σ_0^2 . As σ_0^2 increases, there is more shrinkage to zero as more the spike is relatively more salient.

The behavior of the effect-size estimate under the spike-and-slab specification lead to the following consequences: 1. Estimation is made within the context of a model, and the obtained values are a function of the specification. Model estimates with the spike-and-slab prior show adaptive shrinkage that is useful. 2. Parameter estimation is not in itself more robust to prior settings than the Bayes factor. How robust estimation is to prior settings is a function of model specification.

These two consequences, that the value of estimates and its robustness to prior settings depend critically on the model specification, holds inferences drawn from credible intervals as well. Figure 5C-D show the comparison of credible intervals. For the slab-only specification, the credible intervals maintain a constant width for all observed effect size values. The vertical dashed lines show transition points—it the observed values are more extreme, then the 95% CI does not include zero. The spike-and-slab specification results in different behavior for the credible intervals. For extreme observed values, say those greater than .6 in magnitude, the CIs are quite similar to the slab-only specification. For less extreme values, the spike has influence, and the resulting 95% CI often includes the spike. As a result, the transition points are wider—it takes more extreme observed values to localize the 95% CI away from zero.

The unification through spike-and-slab priors highlights similarities and differences between inference from posterior estimation and inference from Bayes factors as they are commonly used in psychology. The similarities are obvious, both methods are sibling approaches in the Bayes' rule family lineage. They rely similarly on specification of detailed models including models on parameters, and updating follows naturally through Bayes' rule. There are differences as well, and the difference we highlight here is that from

model specification. The recommended methods of inference by estimation rely on broad priors that preclude spikes at set points such as points of invariance. The Bayes factor approaches we have developed in Guan & Vandekerckhove (2015), Rouder et al. (2009), Rouder & Morey (2012) and Rouder, Morey, Speckman, & Province (2012), place point-mass on prespecified, theoretically important values. It is this difference in specification that leads to some of the most salient differences in practice.

Which Model Specification To Use?

A critical question for researchers is then which model specification to use. The answer is that the choice depends on the context of the analysis and the goals of the researcher. As a rule of thumb, it makes sense to consider a point mass on zero when zero is a theoretically meaningful or important quantity of interest. For instance, in the usual testing scenarios, researchers consider the “no-effect” baseline to be qualitatively different than effects. The spike-and-slab model instantiates this difference, and in the process license the theoretically useful abstractions of “effect” and “no effect.” In the context of this goal, of stating evidence for or against effects, it is reasonable and judicious to use a spike-and-slab estimation approach or, equivalently, a Bayes-factor summary of the change in the spike probability. In some cases, perhaps ones where measurement is a main goal and where the zero value has no special meaning, a slab-only approach may be best. Researchers in these measurement contexts, however, should avoid drawing inferences about whether or not there are effects in the data as the model specification does not capture such abstractions. There will be some differences across researchers as to which specification is best in any given context. This differences should be welcomed as they are part of the richness of adding value in psychological science (Rouder, Morey, & Wagenmakers, 2016). In all cases, however, researchers should justify their choices in the context of these goals.

Researchers who consider Bayes factors may worry about their dependence on prior settings especially when compared to estimation with slab-only models. This worry is assuredly overstated, and a bit of common sense provides for a lot of constraint. For example, it seems to us unreasonable to consider prior settings that are too small or too large as researchers generally know that true effect sizes in psychological experiments are neither arbitrarily small or large. A lower limit of σ_0 is perhaps 0.2 as researchers rarely search for effect sizes smaller than this value. Likewise, an upper limit is perhaps 1.0 as the vast majority of effect sizes are certainly smaller than this value. Within these reasonable limits, Bayes factors do vary but not arbitrarily so. We have highlighted the Bayes factor values associated with these limits in Figure 3A as filled circles. Here the Bayes factors differ from 1.7 to 2.8 or by about 40%. This variation is not too substantial, and in both cases the evidence for an effect is marginal. Such variation strikes us as entirely reasonable and part-and-parcel of the normal variation in research Rouder et al. (2016). It is certainly less than other accepted sources such as variation in stimuli, operationalizations, paradigms, subjects, interpretations and the like.

The Potential of Spike-And-Slab Models In Psychology

We think spike-and-slab priors are going to gain popularity as psychologists develop and adopt new analytic techniques, especially in big-data applications. Consider applications in imaging where there are a great many voxels or in behavioral genetics where there are a great many nucleotide markers in a SNP array. It is desirable to consider the activity in any one voxel or the contribution to behavior of any one marker which necessitates the use of models with large numbers of parameters. It is in this context, when there are large numbers of parameters especially relative to the sample size, spike-and-slab priors have become an invaluable computational tool for assessing effects, say which voxels are active or which alleles covary with a behavior. The seminal article for

assessing covariates in this context is George & McCulloch (1993), and recent conceptual and computational advances, say from Scott & Berger (2010) and Ročková & George (2014), make the approach feasible in increasing large big-data contexts.

As an example of big-data applications in psychology, we highlight the recent work of Sanyal & Ferreira (2012) who used spike-and-slab priors for fMRI analysis. These researchers sought to enhance the spatial precision of imaging by improving the spatial smoothing. Typically, researchers smooth the image by passing a Gaussian filter over it. Sanyal and Ferreira instead performed a wavelet decomposition where activation is represented as having a location and a resolution. In this approach there is a separated wavelet coefficient for each resolution and location pairing, and the upshot is a proliferation of coefficients. Sanyal and Ferreira placed a spike-and-slab prior on these coefficients, and used large values of ρ_0 , the prior probability that a coefficient is zero. In analysis, the posterior for many of these coefficients remained dominated by the spike, and could be removed. When the activation was reconstructed from only the coefficients for which there was substantial mass from the slab, the image had improved quality. The resulting smoothing was adaptive—it was more smooth where activation was spatially homogenous (say within structures) and less smooth where activation was spatially heterogeneous (say at boundaries).

Conclusions

In this paper we provide a unification between two competing Bayesian approaches—that based on the estimation of posterior intervals and that based on Bayes factors. A salient difference between these two approaches is in model specification. It is common in estimation approaches to place broad priors over parameters that give no special credence to a zero point. Common Bayes factor approaches, such as that from Rouder and Morey and colleagues (Rouder et al., 2009; Rouder & Morey, 2012; Rouder et

al., 2012; Guan & Vandekerckhove, 2015) are closely related to estimation with a prior that has some point mass at zero. Which model specification a researcher should choose, whether a broad slab or a spike-and-slab, should depend on the context and goals of the analyst.

References

- de Finetti, B. (1974). *Theory of probability* (Vol. 1). New York: John Wiley and Sons.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Quantitative Psychology and Assessment*. Retrieved from 10.3389/fpsyg.2014.00781
- Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236, 119–127.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439-453. Retrieved from <http://psycnet.apa.org/doi/10.1037/a0015251>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman and Hall.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 57-64.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881-889.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. London: Cambridge.
- Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*. Retrieved from <http://www.cidlab.com/prints/guan2015bayesian.pdf>
- James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (p. 361-379).
- Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.

- Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.
- Kruschke, J. K. (2011b). *Doing Bayesian analysis: A tutorial with R and BUGS*. Academic Press.
- Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*, 1242-1249.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187-192.
- Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: Inference*. Cambridge, England: Cambridge University Press.
- Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023-1032.
- Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, -. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0022249615000723>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573-604.
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877-903. Retrieved from <http://dx.doi.org/10.1080/00273171.2012.734737>
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374. Retrieved from <http://dx.doi.org/10.1016/j.jmp.2012.08.001>
- Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological sciencecollabra. *Collabra*, *2*, 6. Retrieved from <http://doi.org/10.1525/collabra.28>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237. Retrieved from <http://dx.doi.org/10.3758/PBR.16.2.225>
- Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*(1201-1208).
- Ročková, V., & George, E. L. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, *109*.
- Sanyal, N., & Ferreira, M. A. R. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *Neuroimage*, *63*, 1519-1531.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587-2619.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response time. *Psychological Methods*, *16*, 44-62.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

Footnotes

¹Perhaps such disagreements should be expected given the contentious history of academic statistics. Even null hypothesis significance testing is a contentious hybrid of Fisherian and Neyman-Pearson schools of thought (Gigerenzer et al., 1989; Lehmann, 1993).

²The Bayes factor between Model \mathcal{M}_1 and \mathcal{M}_0 is

$$B_{10} = \frac{1}{\sqrt{n\sigma_0^2 + 1}} \exp\left(\frac{n^2 d^2}{2(n + 1/\sigma_0^2)}\right) \quad (3)$$

where d is the observed effect size given by \bar{Y}/σ .

³The density of a spike-and-slab model is given by

$$f(\delta) = \rho_0 s(\delta) + (1 - \rho_0) \phi(\delta/\sigma_0),$$

where s is the density of the spike, defined next, ϕ is the density of a standard normal, ρ_0 is the prior mass on the spike, and σ_0^2 is the variance of the slab. The density of the spike, s , is known as a Dirac delta function and derived as follows: Consider a normal density centered at zero with some standard deviation η , denoted $g(\delta) = \phi(\delta/\eta)$. The Dirac delta function, s , is defined as the density in the limit that $\eta \rightarrow 0$:

$$s(\delta) = \lim_{\eta \rightarrow 0} \phi\left(\frac{\delta}{\eta}\right) = \begin{cases} \infty, & \delta = 0, \\ 0, & \text{otherwise.} \end{cases}$$

A wonderful graphical demonstration of this limiting property may be found at https://en.wikipedia.org/wiki/Dirac_delta_function.

⁴The resulting posterior density, $f(\delta|\mathbf{Y})$ is

$$f(\delta|\mathbf{Y}) = \rho_1 s(\delta) + (1 - \rho_1) \phi\left(\frac{\delta - \mu_1}{\sigma_1}\right),$$

where

$$\begin{aligned}\sigma_1^2 &= (n + \sigma_0^{-2})^{-1} \\ \mu_1 &= nd\sigma_1^2 \\ \rho_1 &= \frac{\rho_0}{\rho_0 + (1 - \rho_0)B_{01}},\end{aligned}$$

where d is the observed effect size and B_{01} is the Bayes factor between Model \mathcal{M}_0 and \mathcal{M}_1 .

Figure Captions

Figure 1. A posterior distribution localizes mass providing a view of which values are plausible and implausible. **A.** The posterior localizes the effect away from zero perhaps providing evidence for an effect. **B.** The posterior localizes the effect around zero perhaps providing a lack of evidence for an effect. Dashed lines indicate the 95% credible intervals on posteriors.

figure.1 *Figure 2.* Effects of prior setting σ_0 on posteriors and on inference from credible intervals. **A.** Posterior distributions on effect size δ for $N = 10$ and for a sample effect size of .35. for three settings of σ_0 . **B.** Minimum observed effect sizes needed such that the posterior 95% credible interval excludes zero. The two lines are for sample sizes of 10 and 40, respectively. The results show a robustness to the prior setting of σ_0 .

figure.2 *Figure 3.* The dependence of Bayes factor on prior setting σ_0 . **A.** Bayes factor as a function of σ_0 for $N = 40$ and for an observed effect size of .35. **B.** Minimum observed effect sizes needed such that Bayes factor favors the alternative by 3-to-1. The two lines are for sample sizes of 10 and 40, respectively.

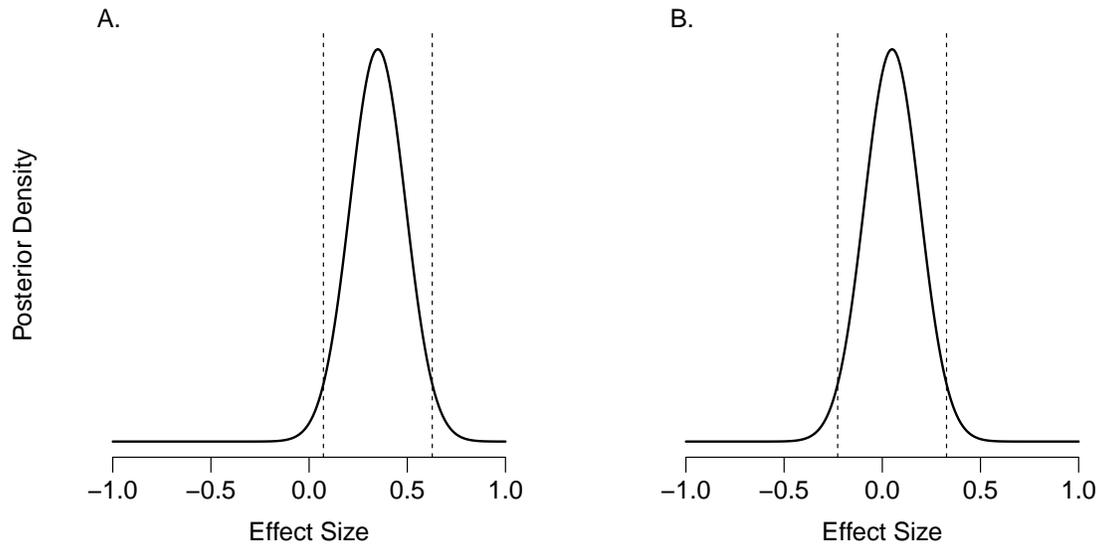
figure.3 *Figure 4.* The spike-and-slab model is a mixture of a spike, shown as an arrow, and slab, shown as the normal curve. **A.** Prior distribution on effect size with half the mass in the spike, and the slab centered around zero. **B-C.** The posterior on effect size δ for observed effect sizes of $d = .1$ and $d = .5$, respectively, for a sample size of 40.

figure.4 *Figure 5.* A comparison of slab-only (\mathcal{M}_1) and spike-and-slab (\mathcal{M}_s) specifications for a moderate sample size of $N = 40$. **A-B:** Posterior mean of δ as a function of d for a few prior settings of σ_0^2 . The light grey line is the diagonal, and the posterior mean of the slab-only model approaches this diagonal as the prior becomes more diffuse. The posterior mean in the spike-and-slab model shows *adaptive shrinkage* where

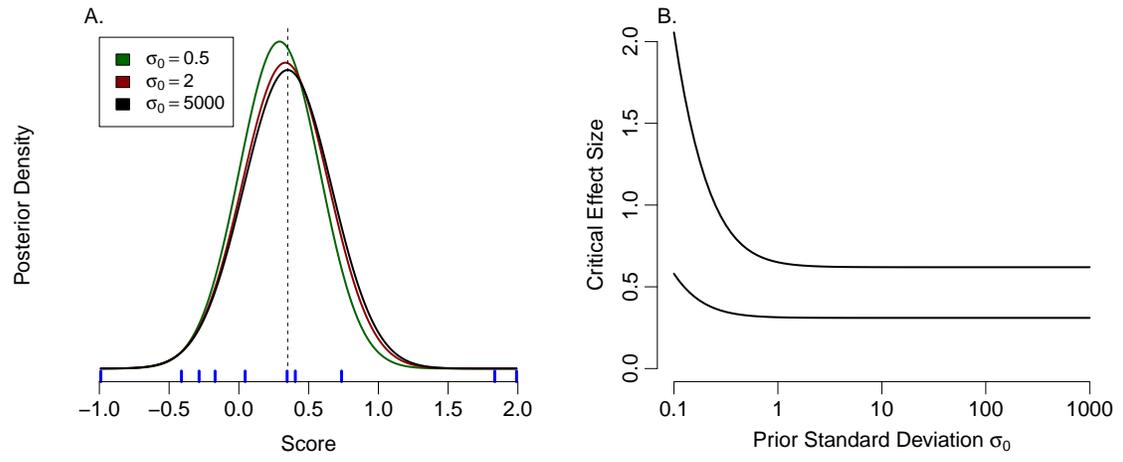
small values observed values result in greatly attenuated estimates. **C-D**: The posterior means with 95% credible intervals. The vertical lines denote transition points—the credible interval does not include zero when the observed effect size is more extreme than these points. The transition points are more extreme for the spike-and-slab specification than the slab-only specification, and this fact is a direct consequence of the point-mass at zero.

figure.5

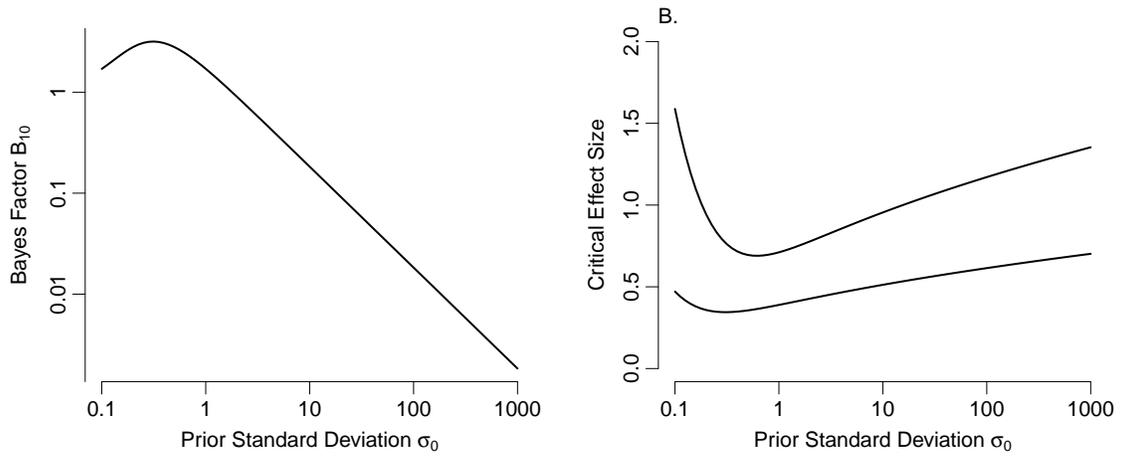
Parameter Estimation and Bayes factors, Figure 1



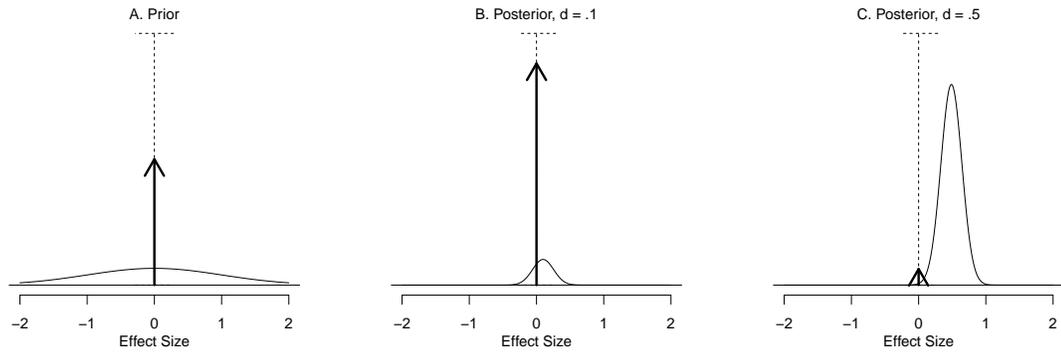
Parameter Estimation and Bayes factors, Figure 2



Parameter Estimation and Bayes factors, Figure 3



Parameter Estimation and Bayes factors, Figure 4



Parameter Estimation and Bayes factors, Figure 5

