

A proposal for informative default priors scaled by the standard error of estimates

March 31, 2021

Abstract

If we have an unbiased estimate of some parameter of interest, then its absolute value is positively biased for the absolute value of the parameter. This bias is large when the signal-to-noise ratio (SNR) is small, and it becomes even larger when we condition on statistical significance; the winner’s curse. This is a frequentist motivation for regularization or “shrinkage”. To determine a suitable amount of shrinkage, we propose to estimate the distribution of the SNR from a large collection or “corpus” of similar studies and use this as a prior distribution. The wider the scope of the corpus, the less informative the prior, but a wider scope does not necessarily result in a more diffuse prior. We show that the estimation of the prior simplifies if we

require that posterior inference is equivariant under linear transformations of the data. We demonstrate our approach with corpora of 86 replication studies from psychology and 178 phase 3 clinical trials. Our suggestion is not intended to be a replacement for a prior based on full information about a particular problem; rather, it represents a familywise choice that should yield better long-term properties than the current default uniform prior, which has led to systematic overestimates of effect sizes and a replication crisis when these inflated estimates have not shown up in later studies.

Keywords: shrinkage, exaggeration ratio, type M error, winner's curse

1 Introduction

Regression modeling plays a central role in the biomedical and social sciences. Linear and generalized linear models, generalized estimating equations, and quantile regression offer great flexibility and are easy to use. When the sample size is not too small, statistical inference can be based on the fact that M-estimates of regression coefficients are approximately normal and unbiased [Stefanski and Boos, 2002]. This yields the familiar frequentist inference in terms of normality-based confidence intervals and p -values, and it also leads to informative Bayesian approaches in which the unbiased estimates form a

likelihood which can be augmented with hierarchical models and other forms of prior information.

If we have an unbiased estimate of some parameter of interest, such as a regression coefficient, then by Jensen’s inequality its absolute value is positively biased for the absolute value of the parameter. This bias is large when the signal-to-noise ratio (SNR) is small, and it becomes even larger when we condition on statistical significance. This is called the winner’s curse or type M error [Gelman and Tuerlinckx, 2000, Ioannidis, 2005, Gelman and Carlin, 2014]. We conclude that noisy estimates must be regularized or partially pooled toward zero. However, the degree of this shrinkage should be carefully considered. Too little shrinkage means that we will systematically overestimate effect sizes, which then later do not replicate. On the other hand, too much shrinkage could lead to missing real discoveries.

From the Bayesian perspective, the right amount of shrinkage depends on the prior. Here, we propose to obtain the relevant prior information from a large collection or “corpus” of similar studies. Such a prior can then be used for default or routine Bayesian inference. The wider the scope of the corpus, the less informative the prior and the more generally applicable. Moreover, a wide scope allows us to include many studies in the corpus so that we can estimate the prior information accurately. Perhaps the most important point we want to make is that a wide scope does not necessarily result in a more

diffuse prior.

In the next section, we will motivate the present paper by discussing in more detail why noisy estimates must be regularized. Then, we argue that we can determine the suitable amount of shrinkage by estimating the distribution of the signal-to-noise ratio (SNR) in a particular area of research. We show that a particular independence assumption will make the estimation easier, and ensures that the posterior inference is unaffected by changes of measurement unit. We also show that depending on the shape of the distribution of the SNR, the amount of shrinkage will be adaptive to the SNR. We demonstrate our approach with two examples. We end the paper with a discussion.

2 Exaggeration and the need to shrink

We will ignore small sample issues by assuming that we have a normally distributed, unbiased estimate b of a regression coefficient β with known standard error s . In other words, conditionally on β and s , b has the normal distribution with mean β and standard deviation s and therefore we have the confidence statement

$$\Pr(b \in [\beta \pm 1.96 s] \mid \beta, s) = 0.95. \quad (1)$$

Our setup may appear to be overly simplistic, but this type of inference is common practice especially concerning regression parameters. And, as noted above, this is also the building block for the standard Bayesian approach; even if one does not care about unbiased estimation per se, independent normally distributed unbiased estimates can be used to construct a likelihood function.

We assume that standard errors are known (i.e. observed without noise), when in reality that is almost never the case. However, it is a very common assumption as most statistical software packages report Wald confidence intervals and associated p -values for log odds ratios (logistic regression), log intensity ratios (Poisson regression) or log hazard ratios (Cox regression). Of course, exact intervals based on the t distribution are reported for linear models, but the difference is already very small in most real-world examples. Our paper is closely related to the field of meta-analysis where standard errors are also usually assumed to be known. For example, the textbook presentation of Bayesian meta-analysis in chapter 5 of the book Bayesian Data Analysis treats the standard errors as known, even though this is an approximation [Gelman et al., 2013].

Since b is unbiased for β , it follows from Jensen's inequality that $|b|$ is positively biased for $|\beta|$. This bias is large when b is noisy, i.e. when the signal-to-noise ratio (SNR) $|\beta|/s$ is small. The bias becomes even larger when we condition on statistical significance, which is called the winner's curse. The relation

between overestimation of the effect size and the SNR has been demonstrated through simulation [Gelman and Carlin, 2014, Ioannidis, 2008], and more recently, the following theorem has been established [van Zwet and Cator, 2021].

Theorem 1. *Suppose b is normally distributed with mean β and standard deviation s . For every $c > 0$, the exaggeration ratio,*

$$\mathbb{E}(|b/\beta| \mid s, \beta, |b| > c)$$

depends on β and s only through the absolute value of the SNR. The exaggeration ratio is always greater than 1. It is decreasing in the absolute value of the SNR and increasing in c .

The exaggeration ratio is also known as the type M error [Gelman and Carlin, 2014].

In Figure 1 we show the extent of the problem by plotting the conditional bias $\mathbb{E}(b - \beta \mid s, \beta, |b/s| > 1.96)$ and exaggeration ratio $\mathbb{E}(|b/\beta| \mid s, \beta, |b/s| > 1.96)$ as a function of the SNR. For the plot of the bias we made the additional assumption that the standard error of the estimate is one.

A partial solution to this overestimation of effect size is to use “weakly informative” priors [Gelman et al., 2008, Greenland and Mansournia, 2015], but then the question arises: how informative should the priors be? The literature on weakly informative priors tends to focus on superior performance compared to noninformative priors. Here we propose to obtain realistic but

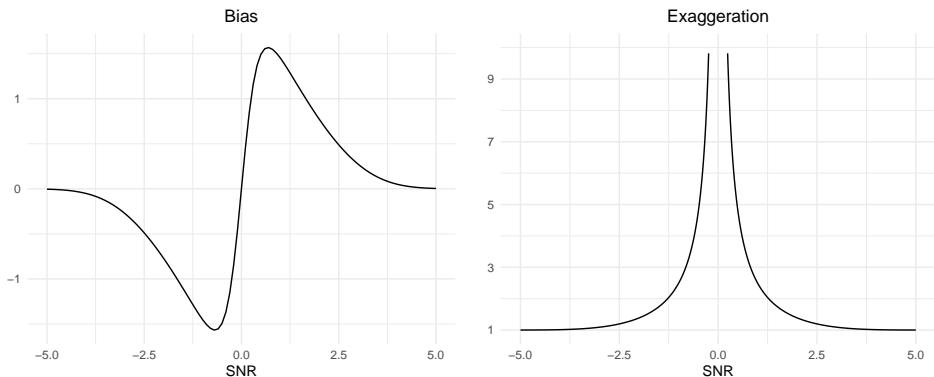


Figure 1: *The bias and exaggeration ratio as a function of the signal-to-noise ratio (SNR), conditional on statistical significance at the 5% level (two-sided).*

general prior information from large collections or “corpora” of similar studies. Such priors can be used for default or routine Bayesian inference. The priors we propose can be narrow and result in a considerable degree of shrinkage.

3 Constructing a default informative prior

3.1 Using a corpus of previous studies

Researchers in the life sciences often believe that they have little or no prior information because their study is unique; nobody has ever studied that particular intervention or exposure in that particular population with that particular outcome under those particular circumstances. We believe that

it is a mistake to think like that. At the highest level of aggregation, just knowing that you are doing another study in the domain of the life sciences represents a lot of information.

It is often possible to be more specific, but that does involve making choices that depend on the details of the study in question. The more we zoom in, the smaller the set of relevant examples becomes. This will make it harder to determine the prior distribution accurately. So, we propose to obtain prior information from large, broad collections of studies.

For our purposes here, we define a *corpus* as a collection of pairs (b_j, s_j) from studies j that are similar in the sense that they meet certain inclusion criteria. An example would be placebo-controlled phase 3 randomized clinical trials (RCTs). If we know only that a particular study meets a set of inclusion criteria—or we are willing to ignore all other features of that study—then we can model that study exchangeably with all others that meet those criteria. The inclusion criteria of the corpus represent exactly the information that we are including in the prior. This implies that if we make the scope wider by removing certain criteria, while keeping the size of the corpus the same, then we will have less information in the prior. For example, a corpus of RCTs is less informative than an equally sized corpus of phase III RCTs because the latter carries the extra information that all the trials are in phase III.

In other words, there is a trade-off between informativeness and general-

ity. However, there is no reason to expect that fewer inclusion criteria would yield a more widely dispersed prior distribution. Indeed, the prior can become less dispersed if by widening the scope, we add mostly studies with small effects. From this point of view, one cannot tell by looking at the prior distribution how much substantive information it represents. This is a different point of view from the usual conviction that high-variance priors carry little information [Kass and Wasserman, 1996]. The current default procedure of a flat prior is appropriate in an unrealistic setting in which all sizes of effects are equally likely.

Obtaining prior information from a corpus with a wide scope has two important practical advantages. First, the wide scope ensures that many studies are eligible so that the prior can be estimated accurately. Second, the resulting prior can be used as a default in a wide range of applications.

3.2 Estimating the prior

Consider three random variables: β , b , and s , with joint density p in the defined population. We assume that b has the normal distribution with mean β and standard deviation s . We observe b and s , and we want to do Bayesian inference about β . Conditionally on s , Bayes' rule states

$$p(\beta | b, s) \propto p(b | \beta, s)p(\beta | s), \quad (2)$$

where $p(b | \beta, s)$ is the conditional density of b given β and s , and $p(\beta | s)$ is the conditional density of β given s . Note that since we are conditioning on s throughout, we do not need to consider its distribution.

To be able to do Bayesian inference about β , our goal is to estimate the probability kernel $p(\beta | s)$ on the basis of a sample of observed pairs (b_j, s_j) . Since we are assuming that $p(b | \beta, s)$ is $\text{normal}(\beta, s)$, it is actually possible to estimate the full joint distribution of (β, b, s) from just the pairs (b_j, s_j) .

Estimating the conditional distribution of β given s is equivalent to estimating the conditional distribution of the signal-to-noise ratio β/s given s . The z value b/s is the sum of β/s and an independent standard normal random variable. So, we can first do some regression modeling to estimate the conditional distribution of b/s conditional on s and then do a deconvolution to obtain the conditional distribution of β/s given s . Finally, we obtain the prior distribution of β given s by scaling.

Suppose we succeed in accurately estimating $p(\beta | s)$ from a large corpus of exchangeable pairs (b_j, s_j) . If we use that estimate as a prior, then the posterior will be approximately calibrated with respect to that corpus. That is, posterior probabilities will represent frequencies across the corpus.

3.3 Independence assumption

Our goal is to develop priors that are widely applicable for routine Bayesian inference. In that context, it is desirable that inference about β does not depend on inconsequential data transformations, such as switching events and non-events in logistic regression, relabeling dummies, changing the unit of measurement of covariates or a numerical outcome. **For example, it should not matter for our substantive conclusions if the data are presented in grams or kilograms.** Mathematically, such a requirement means that,

$$p(\beta \mid b, s) = |c| p(c\beta \mid cb, |c|s), \text{ for all } c \neq 0. \quad (3)$$

Requirement (3) means that posterior inference about β is equivariant under linear transformations of the data. It has the following interpretation in terms of the SNR β/s .

Theorem 2. *Requirement (3) holds if and only if*

1. *s and β/s are independent, and*
2. *the distribution of β/s is symmetric around zero.*

Requirement (3) drastically simplifies the estimation of the probability kernel $p(\beta \mid s)$, because we need only estimate the *marginal* density of β/s . Moreover, we may assume this density is symmetric. We only have to estimate the symmetric marginal density of b/s and then do a deconvolution to

obtain the marginal density of β/s . We can then get the distribution of β given s by simple scaling.

We motivated (3) from a pragmatic point of view by insisting that posterior inference about β should be equivariant under linear transformations to avoid cheating. However, (3) can also be interpreted as an assumption about the joint distribution of the observables b and s . Since b/s is the sum of β/s and an independent standard normal random variable, a trivial consequence of Theorem 2 is

Corollary 1. *Requirement (3) hold if and only if*

- *s and b/s are independent, and*
- *the distribution of b/s is symmetric.*

We can check if these properties hold, at least to reasonable approximation, in any particular corpus. An important necessary condition for the above to hold, is that s and $|b|$ are positively correlated. We argue from an anthropic principle [Gelman, 2018] that it is reasonable to expect such a correlation, as follows. Studies are commonly designed to have just enough power so that effects can just about be estimated from data. Indeed, the goal of sample size calculations (formal or informal) is to balance $|b|$ and s so that the probability that $|b|/s$ exceeds 1.96 is not too large or too small. Hence effects tend to be of the same order of magnitude as standard errors.

This does not preclude that the distribution of the SNR can differ between corpora. For example, some research areas might have larger effects, better measurement devices or more funding opportunities for large studies.

If (3) holds then the prior information about the SNR is all we need for estimating β . Moreover, we have the following equality for the posterior mean

$$\mathbb{E}(\beta | b, s) = s \mathbb{E}(\text{SNR} | z). \quad (4)$$

If (3) does not hold then this equality does not hold either, but we claim that it is still sensible to use the shrinkage estimate $s \mathbb{E}(\text{SNR} | z)$. By conditioning on z , we are using the prior information about the SNR. So, as far as shrinkage is concerned, we are using all the relevant information.

3.4 Adaptivity and consistency with a t prior

The estimate b and its standard error s depend on the sample size, but until now we have suppressed this dependence from our notation. In this section, we will discuss what happens when the sample size increases. Therefore, we will now make the dependence on the sample size explicit.

Suppose we have a normally distributed, unbiased estimator b_n of β with known standard error $s_n = \sigma/\sqrt{n}$, where n is the sample size. If we choose a fixed prior for β , which does not depend on n , then its influence disappears

as the sample size increases in the sense that the posterior distribution of β converges to the likelihood of b_n . In particular, the posterior mean of β converges to b_n and hence is a consistent estimate of β . This is a special case of the well-known Bernstein–von Mises theorem. Here, we are proposing to use a fixed prior for the signal-to-noise ratio β/s_n . Thus, the implied (scaled) prior for β depends on the sample size, and therefore Bernstein–von Mises does not apply.

For example, if we put a normal prior with mean zero and standard deviation τ on β/s_n , then the posterior mean for β is $\frac{\tau^2}{\tau^2+1}b_n$. Evidently, this is an inconsistent estimate of β , unless β happens to be zero.

Fortunately, by choosing a prior with flatter tails than the normal, it is possible to have a fixed prior on β/s_n and still have the posterior distribution of β converging to the likelihood of b_n . The following theorem is a special case of a result due to Dawid in the context of Bayesian outlier detection [Dawid, 1973].

Theorem 3. *Suppose we have a normally distributed, unbiased estimate b_n of β with known standard error $s_n = \sigma/\sqrt{n}$, where n is the sample size, and suppose β is assigned a sample-size-dependent $t_\nu(0, s_n)$ prior distribution. Then, as long as the true β is not equal to zero, the limiting posterior distribution of $(\beta - b_n)/s_n$ is standard normal.*

The point is that the t distribution has a much flatter tail than the normal distribution. As n becomes large, the likelihood of b_n will concentrate around the true and nonzero β . Meanwhile, the prior, by construction, becomes increasingly narrow and is centered around 0. Thus the overlap with the normal likelihood will be in a region where the prior is almost completely flat and hence the posterior will converge to the likelihood. In other words, as n grows and the z -value b_n/s_n becomes large, the shrinkage disappears. In that sense, the amount of shrinkage adapts to the signal-to-noise ratio.

Dawid's theorem is actually more general and provides sufficient conditions for the tail behavior of the prior. There is an extensive literature about heavy-tailed priors which is reviewed by O'Hagan and Pericchi [O'Hagan and Pericchi, 2012].

3.5 Mixture of normals prior

Above we established that using a t prior distribution for β/s yields a consistent estimator. In practice, we prefer to use a finite mixture of zero-mean normal distributions, with a density of the form

$$f(x) = \sum_{i=1}^k p_i \varphi(x/\tau_i) / \tau_i, \quad (5)$$

where φ denotes the standard normal density, τ_i are the standard deviations of the k mixture components and the non-negative mixture proportions p_i

add up to one. This model has two advantages. First, all calculations can be done explicitly, which is fast and can give us insight. The mathematical details are not difficult, and we describe them in the online supplement. More importantly, a mixture of zero-mean normal distributions is a very flexible model. Already with just two components, we can separately fit the central part and the tails of the distribution of β/s . As it turns out, a mixture of two components provides a reasonably good fit in our examples, so this is what we used there.

We estimate the mixture proportions and the variances of the mixture components by maximum likelihood. For the analyses of the next section, we have used the R package “flexmix” [Leisch, 2004], which implements the EM algorithm. Under our assumptions, a z -value is the sum of the SNR and standard normal noise. Therefore the variances of the mixture components of the distribution of the z -values must be at least one. So, to estimate the mixture distribution by maximum likelihood we must actually solve a constrained optimization problem. When we estimate the distribution of the z -values in the examples of the next section, it turns out that the constraint is not active because the likelihood is maximized with all variances greater than one.

The tails of a mixture of Gaussians are not heavy enough to meet the requirements of Dawid’s theorem, and therefore we do not get formal con-

sistency. However, this is not a major practical concern. Sample sizes never actually go to infinity, and if one of the components has a large standard deviation, then the tails of the mixture will be flat enough in the sense that there will be little shrinkage when the observed z -value is large. As with many statistical models (for example, logistic vs. probit regression), what is most important is not the exact functional form but rather that the model has enough flexibility that we can learn from data.

4 Example using corpora in psychology and medicine

We will illustrate the ideas of this paper with two example corpora, one from psychology and one from medicine.

To obtain reliable prior information, the reported effects in our corpus must be a fair sample of the population of effects within the scope. It is well-known that for various reasons (publication bias, file drawer effect, researcher degrees of freedom, fishing, forking paths, etc.) reported effects tend to be inflated [Ioannidis, 2005, Rothstein et al., 2006, Ioannidis, 2008, Button et al., 2013, Gelman and Loken, 2014, Collaboration, 2015]. Here, we consider two special cases where the risk of publication bias is low so

that we expect to find a reasonably honest sample of effects.

4.1 Open Science Collaboration study on reproducibility in psychology

To assess the reproducibility of psychological science, the Open Science Collaboration (OSC) selected 100 studies from three leading journals of the American Psychological Association, and replicated them [Collaboration, 2015]. They chose the journals Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition. According to the authors of the replication study, the first journal is a premier outlet for all psychology research; the second and third are leading disciplinary-specific journals for social psychology and cognitive psychology, respectively. The studies were selected in a quasi-random way to balance two competing goals: “minimizing selection bias by having only a small set of articles available at a time and matching studies with replication teams’ interests, resources, and expertise.” The data are publicly available at <https://osf.io/fgjvw/>.

For our corpus, we use only the replication studies from the OSC project. Since the regression parameters and their standard errors are not available in the data set, we transformed the two-sided p -values of the replication studies

to absolute z -values by

$$|z| = -\Phi^{-1}(p\text{-value}/2).$$

Under requirement (3) the absolute z -values are sufficient to estimate the prior. Excluding the F -tests, we have 86 absolute z -values. We show their distribution in Figure 2. Since we have only the absolute values, we symmetrized the histogram, that is, we used the R command `hist(c(-z,z))`.

We used maximum likelihood (via the EM algorithm) to estimate the distribution of the z -values as a two-component mixture of zero-mean normals. The R code is available in the online supplement. We find a mixture with standard deviations 1.2 and 4.1 and mixture proportions 0.57 and 0.43, respectively. We show this distribution in Figure 2.

Now recall that the SNR is the sum of the z -value and independent standard normal noise. Therefore, we can obtain the distribution of the SNR by deconvolution. This is quite complicated in general, but it is trivial when using normal distributions. The distribution of the SNR has the same mixing proportions as the distribution of the z -values, and the standard deviations are simply $\tau_1 = \sqrt{1.2^2 - 1} = 0.7$ and $\tau_2 = \sqrt{4.1^2 - 1} = 4.0$. We show this distribution in the left panel of Figure 3.

Recalling (4), we can use the distribution of the SNR to obtain a shrinkage estimator of the parameter β from the observed z -value as $\hat{\beta} = s\mathbb{E}(\text{SNR} | z)$.

Computing $\hat{\beta}$ is not difficult, and we provide a few lines of R code in the appendix. We refer to the online supplement for the mathematical details. We can now define the *shrinkage factor* as

$$\frac{b}{\hat{\beta}} = \frac{z}{\mathbb{E}(\text{SNR} | z)}. \quad (6)$$

We show the shrinkage factor in the right panel of Figure 3 as a function of the z -value.

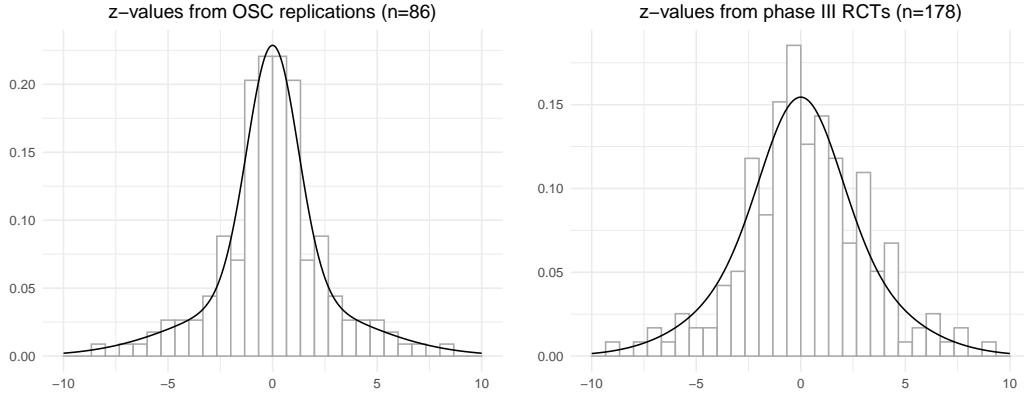


Figure 2: *The observed z -values of the replication studies in psychology from the Open Science Collaboration (OSC), and the Phase III randomized controlled clinical trials (RCTs) in medicine from the Cochrane collaboration. For the OSC study we show the symmetrized histogram (see section 4.1). For both datasets we also show the fitted mixtures of two zero-centered normals.*

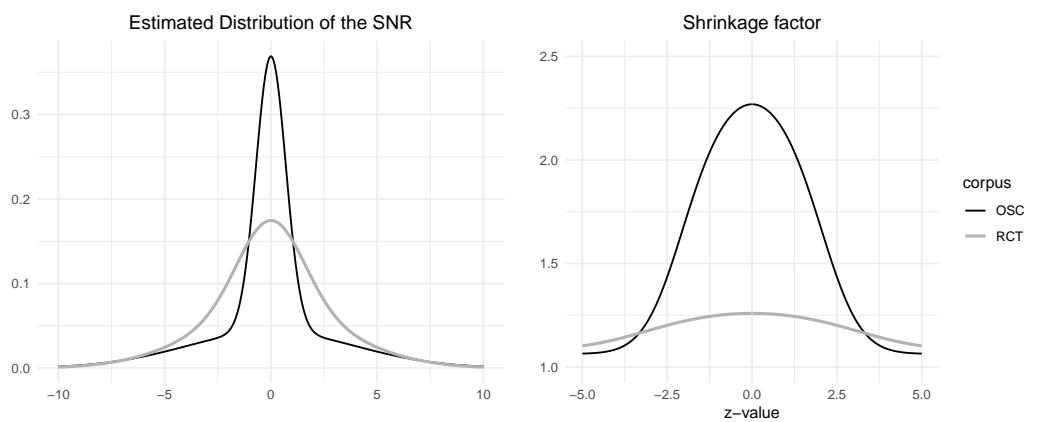


Figure 3: *Left panel:* the estimated distribution of the SNR for the replication studies in psychology from the Open Science Collaboration (OSC), and the Phase III randomized controlled clinical trials (RCTs) in medicine from the Cochrane collaboration. *Right panel:* The shrinkage factor—the amount by which the raw estimate is divided to yield the Bayes estimate—as a function of the z-score of a new study, for each of our two classes of problems. A shrinkage factor of 1 corresponds to no shrinkage.

4.2 Cochrane phase 3 placebo controlled clinical trials

The Cochrane Database of Systematic Reviews (CDSR) is the leading journal and database for systematic reviews in health care. All z -values from primary studies up to 2018 have been derived, and made public [Schwab, 2020]. We decided to focus on phase III randomized, placebo controlled clinical trials (RCTs). Such trials involve large groups of patients and are aimed at being the definitive assessment of the effectiveness of a particular treatment. They represent large investments and are typically the culmination of many years of research and development. It is therefore quite unlikely that their results go unpublished. Moreover, the Cochrane Collaboration makes every effort to include all relevant studies for their systematic reviews – even unpublished studies.

We selected studies from the CDSR where either the title or Cochrane’s methods description contained the terms “phase 3” or “phase III.” Next, we selected all comparisons for efficacy against a placebo. For each study, we selected only a single z -value where we tried to obtain the effect of primary interest. We were left with 178 z -values. We show their distribution in Figure 2.

We estimated the distribution of the z -values as a two-component mixture of zero-mean normals with standard deviations 1.8 and 3.7 and mixture

proportions 0.42 and 0.58, respectively. We show this distribution in Figure 2. Again, we can obtain the distribution of the SNR by deconvolution. This distribution has the same mixing proportions as the distribution of the z -values, and the standard deviations are $\tau_1 = \sqrt{1.8^2 - 1} = 1.5$ and $\tau_2 = \sqrt{3.7^2 - 1} = 3.5$. We show this distribution in the left panel of Figure 3. We show the shrinkage factor in the right panel of Figure 3.

4.3 A typical application

To illustrate the implications of our proposal, we will now work through a small numerical example. All computations in this section are in the online supplemental document. Consider a study in the field of psychology where we compare two groups on some binary outcome. Suppose the estimated log odds ratio is $b = 0.3$ with a standard error of $s = 0.2$. The Wald 95% confidence interval is from -0.1 to 0.7.

Suppose we are willing to view the study of interest as exchangeable with the studies in the OSC corpus. We provide an R function in the boxed listing to compute the posterior distribution of the “true” log odds ratio β . The function call is: `posterior(b=0.3,s=0.2,p=c(0.57,0.43),tau=c(0.7,4.0))`. The posterior distribution is again a mixture of two normal components with mixing probabilities 0.7 and 0.3, means 0.1 and 0.28, and standard deviations

```

posterior <- function(b,s,p,tau){
  z <- b/s
  tau2 <- tau^2
  q <- p*dnorm(z,0,sqrt(tau2+1))
  q <- q/sum(q)                      # conditional mixing probs
  pm <- b*tau2/(tau2+1)              # conditional means
  pv <- s^2*tau2/(tau2+1)            # conditional variances
  ps <- sqrt(pv)                     # conditional std devs
  data.frame(q,pm,ps)
}

```

Listing 1: R code to compute the posterior of β given (b, s) when the prior distribution of the SNR is a mixture of zero-mean Gaussians with mixing probabilities $p = (p_1, p_2, \dots, p_k)$ and standard deviations $\tau = (\tau_1, \tau_2, \dots, \tau_k)$.

0.11 and 0.19. Hence, the posterior mean of β is $\hat{\beta} = 0.7 \times 0.1 + 0.3 \times 0.28 = 0.15$. The shrinkage factor is $b/\hat{\beta} = 0.3/0.15 = 2$. The 95% credible interval is -0.12 to 0.55, and the posterior probability that β is actually negative is 0.16.

4.4 Remarks

The two datasets we have studied yielded very different results as can be seen in Figures 2 and 3. Unsurprisingly, the signal-to-noise ratio in psychology research tends to be much smaller than in phase 3 clinical studies. The latter usually involve very considerable investments and hence may be expected to have high statistical power. Consequently, it seems that one should apply much stronger shrinkage to results from psychological research than from

phase 3 clinical studies, especially when the observed (absolute) z -value is small.

An important conclusion of the OSC reproducibility study was that on average the effect size of the replication effects was half the magnitude of the effect size of the original effects [Collaboration, 2015]. This roughly agrees with our analysis. Using *only* the results of the replication studies, we found that shrinkage by a factor of about 1.5 to 2 is typically in order.

A more extensive analysis of the entire Cochrane database is reported elsewhere [van Zwet et al., 2020].

5 Discussion

5.1 The value of default informative priors

Nearly forty years ago, Rubin wrote [Rubin, 1984]:

Another reason for the applied statistician to care about Bayesian inference is that consumers of statistical answers, at least interval estimates, commonly interpret them as probability statements about the possible values of parameters. Consequently, the answers statisticians provide to consumers should be capable of being interpreted as approximate Bayesian statements.

The present paper is an attempt to do just that. The confidence interval (1) describes the long-run coverage performance of the random interval $[b - 1.96 s, b + 1.96 s]$. The statement does not hold conditionally on the data, but it is often mistakenly interpreted “Bayesianly” as

$$\Pr(\beta \in [b \pm 1.96 s] \mid b, s) = 0.95. \quad (7)$$

where β is viewed as a random variable, and we condition on the data pair (b, s) . We refer to Greenland et al. [Greenland et al., 2016] for a discussion of this misinterpretation. Statement (7) is arguably more relevant than (1) because it refers to the data at hand, rather than the procedure being used. This may explain, at least in part, the pervasiveness of the misinterpretation; it is what researchers want to know.

The Bayesian statement (7) is only valid if β has the (improper) uniform or “flat” prior distribution. The matching property of (1) and (7) has lead many to consider the uniform prior to be an objective or noninformative prior [Ghosh et al., 2011]. Many other criteria have been proposed which a priori might be considered to be objective [Kass and Wasserman, 1996], but in the normal location model with known standard deviation they all yield the (improper) uniform distribution as the unique objective prior. So, we find that the flat prior is used for Bayesian inference about regression coefficients in two distinct situations: explicitly with the goal of objective Bayesian inference

and implicitly whenever the confidence interval for a regression coefficient is interpreted as a credibility interval.

The goal of using a noninformative prior is to be impartial or objective by minimizing the influence of the prior on the posterior, see Berger [Berger, 2006] but also Gelman and Hennig [Gelman and Hennig, 2017]. However, this influence depends on which aspect of the posterior we are considering. The flat prior is actually very informative both for the magnitude and the sign of β . This is just a consequence of the fact that a diffuse prior favors large absolute values. In fact, use of the flat prior results in overestimation of the magnitude of β and exaggerated evidence about its sign [van Zwet, 2019]: type M (magnitude) and type S (sign) errors [Gelman and Carlin, 2014]. We thus echo the classical Bayesian literature in concluding that “noninformative prior information” is a contradiction in terms. The flat prior carries information just like any other; it represents the assumption that the effect is likely to be large. This is often not true. Indeed, the signal-to-noise ratio β/s is often very low and then it is necessary to shrink the unbiased estimate. Failure to do so by inappropriately using the flat prior causes overestimation of effects and subsequent failure to replicate them.

Some degree of shrinkage is achieved by using weakly informative priors. This can provide good results in many situations [Gelman et al., 2008, Greenland and Mansournia, 2015], but can still lead to undershrinkage and

positively biased effect size estimates. Here we propose to use prior information estimated from a large corpus of similar studies, under the constraint that we are modeling effect size in standard error units. By using a wide scope, we can ensure that little information is required that is specific to the study at hand. A wide scope also means that we can include many studies so that the prior can be estimated accurately. Finally, a wide scope means that the prior information is applicable to many studies. A wide scope does *not* imply that the prior will be wide in the sense of having high variance.

If we succeed in accurately estimating the prior information from a large corpus then the resulting posterior inferences will be approximately calibrated with respect to that corpus. That is, posterior probabilities will represent frequencies across the corpus. It is important to distinguish this frequentist Bayesian between-studies perspective from a more typically Bayesian within-study framework view where posterior probabilities represent a study-specific model.

If our corpus-based prior distributions are to be used for default or routine Bayesian inference, then those inferences should not depend on linear data transformations such as a change of the unit of measurement. Requiring our inference to be equivariant under linear transformations of the data greatly simplifies estimation of the prior (Theorem 2). Under this requirement, we only need to estimate the symmetric, marginal distribution of the observed

z -values.

To use a corpus-based prior, one only needs to combine it with the (approximately) unbiased, normally distributed estimate of the parameter of interest and its standard error. This is a great advantage, because it allows anyone to perform a quick Bayesian re-analysis of a standard frequentist result. No need to wait for the author to do that!

5.2 Limitations of our recommended approach

We have assumed throughout that the standard errors of estimates are known, while that is typically not realistic. We argued in section 2 that this is a very common assumption in meta-analysis which is quite harmless when sample sizes are relatively large. Here we want to stress that the assumption is really only necessary for constructing the prior. The subsequent Bayesian inference can proceed without it.

The main difficulty of the method described in the present paper is the need to compile an honest corpus that is not affected by publication bias, file drawer effect, researcher degrees of freedom, fishing, forking paths, etc. Promising sources are replication studies, registered reports or careful meta-analyses that make an effort to include also unpublished studies. In section 4.2 we used only a small part of the Cochrane data [Schwab, 2020]. The

data are much more extensive and can be broken down by medical specialty, purpose (efficacy or safety), type of outcome and more, to yield relatively specific priors. Work along these lines is underway, see also van Zwet et al. [van Zwet et al., 2020].

A second caveat is our pragmatic requirement (3) that our inference should be equivariant under linear transformations of the data. This requirement is important to ensure that it is not possible to manipulate the conclusions of a study by a change of measurement unit or by comparing group B to A instead of A to B. This requirement implies that b/s and s are independent and that the distribution of b/s is symmetric. Those properties may or may not be reasonable in a particular corpus.

Our recommended approach makes use of the Bayesian formalism but is not fully Bayesian in that it does not correspond to any joint distribution of parameters and data. Our choice of prior is improper, not in the sense of having no finite integral but in that it depends on the data (through the sample size n), which is not allowed in Bayesian inference. For any particular dataset, the prior is proper, but the resulting inference violates Bayes' rule as new data come in. For example, suppose an experiment with $n = 100$ is analyzed using the methods described in the present article, and then the researcher goes and performs the study on 300 more people drawn from the same population. We can treat this as new data and do Bayesian updating

using the posterior from the experiment just performed as our prior for the analysis of the 300 new people, or we can consider the data as one experiment and go back to the default prior, this time scaled to $n = 400$. Because of the scaling of the prior, these two inferences will differ.

Arguably, however, some incoherence is appropriate for any default prior for a continuous parameter. An informative prior for any regression coefficient will require some scaling [Gelman et al., 2008], and if this is not based on the data it would require an equivalent restriction to some class of appropriately-scaled problems. The prior we have proposed here is unusual in that it is scaled to sample size, but this can be seen as a sort of rationalized version of current statistical practice which is to judge the plausibility of claims based on their t ratio, the number of standard errors the estimate is from zero.

The availability of a corpus-based prior does not preclude using more specific prior information where available. This can be considered as an implicit restriction of the corpus to a more relevant set of problems.

5.3 A default default prior?

In this article we have proposed a method for constructing a default prior for a class of problems by fitting a wide-tailed distribution (t or mixture

of normals) to data from a relevant corpus of careful studies. But what about a truly default prior, to be applied in new problems, or settings where no reliable corpus is available, or for use in general-purpose software? In this case we could see the virtue of a choice such as the standard Cauchy distribution, which does a lot of shrinkage for noisy estimates but approaches the classical limit as the precision of the estimate increases, as illustrated in Figure 4. We note that the standard Cauchy distribution is the same as the standard t distribution with one degree of freedom, which we denote $t_1(0, 1)$. Other members of the t family could also be considered as priors. There is no magic about this choice, and it will be appropriate only to the extent that this prior reflects the distribution of underlying effects. That said, we believe that this sort of standard-error-scaled prior can be a useful starting point in many settings.

References

- [Berger, 2006] Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402.
- [Button et al., 2013] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power

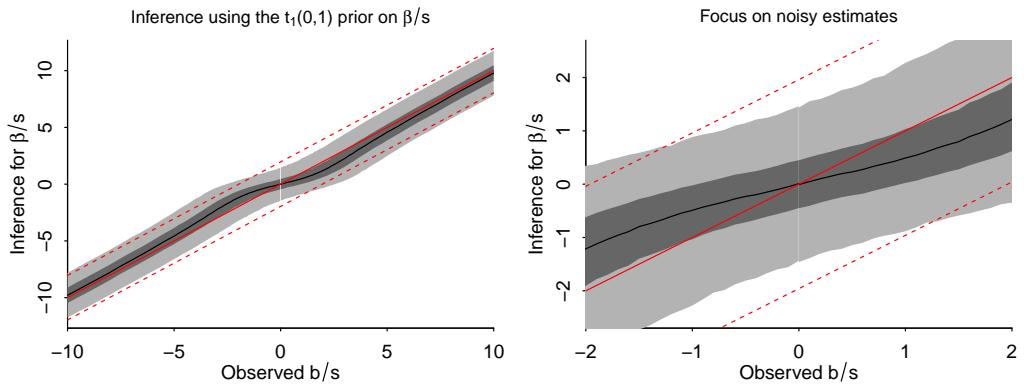


Figure 4: *Inference given a default standard Cauchy or $t_1(0, 1)$ prior on b/s . Line and shaded bands show posterior median, 50%, and 95% intervals. Red line and dotted lines show classical (or flat-prior) estimate and 95% interval. When the estimate is noisy (less than 2 standard errors away from zero, estimates are shrunk about halfway toward zero; as the precision of the estimate becomes higher, the inferences approach the classical limit.*

failure: Why small sample size undermines the reliability of neuroscience.

Nature Reviews Neuroscience, 14(5):365.

[Collaboration, 2015] Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

[Dawid, 1973] Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, 60(3):664–667.

[Gelman, 2018] Gelman, A. (2018). The anthropic principle in statistics. *Statistical Modeling, Causal Inference, and Social Science*. <https://statmodeling.stat.columbia.edu/2018/05/23/anthropic-principle-statistics/>.

[Gelman and Carlin, 2014] Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

[Gelman and Hennig, 2017] Gelman, A. and Hennig, C. (2017). Beyond subjective and objective in statistics (with discussion). *Journal of the Royal Statistical Society, Series A*.

[Gelman et al., 2008] Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383.

[Gelman and Loken, 2014] Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102:460–465.

[Gelman and Tuerlinckx, 2000] Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390.

[Ghosh et al., 2011] Ghosh, M. et al. (2011). Objective priors: An introduction for frequentists. *Statistical Science*, 26(2):187–202.

[Greenland and Mansournia, 2015] Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34(23):3133–3143.

[Greenland et al., 2016] Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350.

- [Ioannidis, 2005] Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2):218–228.
- [Ioannidis, 2008] Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.
- [Kass and Wasserman, 1996] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.
- [Leisch, 2004] Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18. <http://www.jstatsoft.org/v11/i08/>.
- [O’Hagan and Pericchi, 2012] O’Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, 26(4):372–401.
- [Rothstein et al., 2006] Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2006). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Wiley.

[Rubin, 1984] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4):1151–1172.

[Schwab, 2020] Schwab, S. (2020). Re-estimating 400,000 treatment effects from intervention studies in the cochrane database of systematic reviews [data set]. *Open Science Framework*. <https://doi.org/10.17605/OSF.IO/XJV9G>.

[Stefanski and Boos, 2002] Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *American Statistician*, 56(1):29–38.

[van Zwet, 2019] van Zwet, E. W. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research*, 28:3799–3807.

[van Zwet and Cator, 2021] van Zwet, E. W. and Cator, E. A. (2021). The winner’s curse and the need to shrink. *Statistica Neerlandica*, pages 1–15.

[van Zwet et al., 2020] van Zwet, E. W., Schwab, S., and Senn, S. J. (2020). The statistical properties of RCTs. <http://arxiv.org/abs/2011.15004/>.