Author response to reviews of

Selecting relevant moderators with Bayesian regularized meta-regression

masked
submitted to *Research Synthesis Methods*

[RC] Reviewer comment Manuscript text

Dear Dr. Pigott,

Thank you for considering our manuscript for publication in *Research Synthesis Methods*. We appreciate the thorough and critical comments you solicited, which helped strengthen the argumentation, clarify the structure, and add nuance. We have attempted to address all comments as thoroughly as possible. All changes are explained below. We hope that you will find the quality of the work to be sufficiently improved.

Yours sincerely,

the authors

Associate Editor

RC: **The authors introduce their implementation of a regularisation approach for estimation and variable selection in meta-regression, which seems a timely contribution given the danger of overfitting. Some more detail or elaboration would be helpful in some places.**

AR: In addressing all comments below, we have added detail and elaboration in many places. We hope that this addresses this comment.

RC: **please consider whether the title is appropriate, or whether (something like) "Selecting..." or "Selection of..." might be better.**

AR: We appreciate the suggestion and have changed the title to:

Selecting relevant moderators with Bayesian regularized meta-regression

RC: **in abstract and Introduction, please consider being more concrete here: it might be good to mention here already that this is about penalisation/LASSO/horseshoe priors in order to give the reader a better idea.**

AR: To address this comment and others, we have restructured the Introduction to "come to the point" much sooner. We still provide the same background information in a more abbreviated form at a later point. With regard to the specific request to mention penalisation/LASSO/horseshoe priors in the Abstract and Introduction, we have made the following changes. The Abstract now reads:

> we introduce Bayesian Regularized Meta-Analysis (BRMA), which selects relevant moderators from a larger set of candidates by shrinking small regression coefficients towards zero with regularizing (LASSO or horseshoe) priors.

The Introduction now states:

> The present paper introduces *Bayesian Regularized Meta-Regression* (BRMA), an algorithm that uses Bayesian regularizing priors to perform variable selection in meta-analysis. Regularizing priors assign a high probability density to near-zero values, which shrinks small regression coefficients towards zero, thus resulting in a sparse solution. This manuscript discusses two shrinkage priors, the LASSO and horseshoe prior.

RC: on p.16 (Section "design factors"): what were the total numbers of variables (or "noise moderators")?

AR: We see now that we did not report clearly enough how many design factors there were and what they were. Apparently, it was also not sufficiently clear that "noise moderators" are not the same as "design factors". To address this comment and avoid confusion, we have rewritten this paragraph as follows:

> To examine performance in a range of realistic meta-analysis scenarios, seven design factors were manipulated: First, we manipulated the number of studies in the training data $k \in (20, 40, 100)$. Second, the average within-study sample size $\bar{n} \in (40, 80, 160)$. Third, true effect sizes were simulated according to two models: one with a linear effect of one moderator, $T_i = \beta x_{1i} + \epsilon_i$, and one with a non-linear (cubic) effect of one moderator, $T_i = \beta x_{1i} + \beta x_{1i}^2 + \beta x_{1i}^3 + \epsilon_i$, where $\epsilon_i \sim N(0, \tau^2)$. As both BRMA and RMA assume linear effects, simulating data from a non-linear model allows us to examine how robust the different methods are to violations of this assumption. The fourth design factor was the population effect size β in the aforementioned models, with $\beta \in (0, .2, .5, .8)$. Fifth, we manipulated the residual heterogeneity τ^2 in the aforementioned models, with $\tau^2 \in (.01, .04, .1)$. According to a review of 705 published psychological meta-analyses (Van Erp et al., 2017), these values of τ^2 fall within the range observed in practice. Sixth, we varied the number of moderators not associated with the effect size $M \in (1, 2, 5)$. These are the moderators that ought to be shrunk to zero by BRMA. Note that the total number of moderators is $M + 1$, as one moderator is used to compute the true effect size (see the third design factor). Finally, moderator variables were simulated as skewed normal moderators, with scale parameter $\omega \in (0, 2, 10)$, where $\omega = 0$ corresponds to the standard normal distribution.

RC: regarding the 3rd comment by Reviewer 1: in the context of variable selection, the so-called "penalized complexity priors" (here: exponential priors for the heterogeneity standard deviation) might play a role.

AR: To address the 3d comment by Reviewer 2, we have included the following paragraph in the revision:

The choice of prior distributions is an important decision in any Bayesian analysis. This also applies to the heterogeneity parameters. In the case of random effects meta-regression, the only heterogeneity parameter is the between-studies variance, τ^2. In the case of three-level multilevel meta-regression, there is a within-study and between-studies variance.

A crucial challenge with heterogeneity parameters in meta-regression is that the number of observations at the within- and between-study level is often small. This can result in poor model convergence (**roverWeaklyInformativePrior2021?**), or boundary estimates at zero (**chungAvoidingZeroBetweenstudy2013?**). A well-known advantage of Bayesian meta-analysis is that it can overcome these challenges by using weakly informative priors, which guide the estimator towards plausible values for the heterogeneity parameters. There is less consensus, however, about which priors are best for this purpose (**roverWeaklyInformativePrior2021?**). BRMA uses a prior that was specifically developed for multilevel heterogeneity parameters (**gelmanPriorDistributionsVariance2006?**): the half-Student's t distribution with large variance, $df = 3, scale = 2.5$. Note that other suitable weakly informative priors have been discussed in the literature, such as the Wishart prior (**chungWeaklyInformativePrior2015?**). There has also been increasing interest in the use of informative priors for heterogeneity parameters (**thompsonGroupspecificPriorDistribution2020?**). Informative priors incorporate substantive domain knowledge about likely values for the parameter. Informative priors are different from weakly informative priors, which merely guide the estimator towards possible values (e.g., by excluding negative values for the variance), or restrict it to plausible values to aid model convergence. BRMA takes a pragmatic approach to Bayesian analysis, using weakly informative priors to aid convergence for heterogeneity parameters, and regularizing priors to perform variable selection for regression coefficients. The use of informative priors is out of scope for BRMA. If researchers do wish to construct alternative prior specifications, they may want to develop a custom model in `rstan` instead (Stan Development Team, 2022).

1. Reviewer 1

RC: The manuscript provides an interesting approach based on regularization to selection of covariates in meta-regression, when the number of covariates is relevant. The Authors focus on a Bayesian strategy. Although I found the topic interesting, the study needs some additional investigations and the text needs to be substantially modified in order to provide a clear and exhaustive picture of the proposal. My comments and questions are listed below.

AR: We thank the Reviewer for their constructive comments, which substantially helped improve the quality of the manuscript. We have tried to address them all in turn, as detailed below.

RC: The introduction is too long in terms of explanation of the meta-regression and selection of covariates problems, with many repetitions. For example, end of page 4 and beginning of page 5 include many repetitions about between-study heterogeneity. Other parts, instead, require a substantially deeper explanation.

AR: To address this comment and others, we have restructured the Introduction to "come to the point" much sooner. We still provide the same background information in a more abbreviated form at a later point. We have elaborated on all the points that were raised by the Reviewers (see our responses to the other comments).

RC: Which theories useful for variable selection do you refer to in page 5 ? The description is extremely vague.

AR: We have attempted to clarify this section by invoking an additional reference and rewriting it as follows:

> One way to perform variable selection is by relying on theory, and selecting only moderators that should theoretically have an impact on effect size. An important caveat is that theories that describe phenomena at the level of individual units of analysis do not necessarily generalize to the study level. Using such theories for variable selection amounts to committing the ecological fallacy: generalizing inferences across levels of analysis (Jargowsky, 2004). The implications of the ecological fallacy for interpreting the *results* of meta-regression has been discussed elsewhere (**bakerUnderstandingHeterogeneityMetaanalysis2009?**): For example, it is possible for meta-regression to find a significant positive effect of average sample age on the effect of a treatment for cancer, even if age is unrelated with treatment effect within each study. The opposite problem holds when using theory on the individual level of analysis to select study-level moderators: If age does predict treatment effect within each study, that does not mean that average sample age will be a relevant moderator in meta-regression. For an example of theory at the study level of analysis, consider the *decline effect*: effect sizes in a given tranche of the literature tend to diminish over time (Schooler, 2011). When, by coincidence, a large effect is found, it initially draws attention from the research community. Subsequent replications tend to find smaller effects due to regression to the mean. Based on the decline effect, we might hypothesize "year of publication" to be a relevant moderator of study effect sizes. Few theories about the drivers of between-study heterogeneity exist, however, and until they are developed, theory has limited utility for variable selection.

RC: The topic of the study is variable selection from a Bayesian point of view: but the description is so short. Just a few lines about the novelty of the paper in page 6, with no clear description of the proposal and with the reference to pema package without even the explanation of the name of the package.

AR: We have rewritten the

Work in progress

RC: Statistical underpinnings: Again, I found a lot of repetitions in the text.

AR: **Work in progress**

RC: Probably, in (1) you mean θ and not Θ? As you have θ int the text. Why do you use theta and not beta, as you do for the meta-regression model in (7)? The notation would be simplified and would be more consistent.

AR: **Work in progress**

RC: The sentence in lines 132-133 is not clear.

AR: **Work in progress**

RC: line 147: probably the explanation of the error term should be anticipated where the error term appears for the first time.

AR: **Work in progress**

RC: line 152: "to estimate this model " has a preferable statistical sound than "to solve this model"

AR: Work in progress

RC: lines 163- 166 describe some features of regularization. . . .but regularization is introduced later, from line 167.

AR: Work in progress

RC: Regularized regression + Bayesian estimation I found both the sections very short and lacking of many details.

AR: Work in progress

RC: what about the existing literature using lasso for meta-regression?

AR: Work in progress

RC: In addition, I expected some references about the theory of LASSO given the relevance of the instrument in the literature.

AR: Work in progress

RC: line 183-184: "other penalties exist": ok, so discuss about, indicate advantages and disadvantages of the other solutions; why do you not focus on them in the study? If you do not, just explain why.

AR: Work in progress

RC: Bayesian estimation: I think that some additional methodological details and additional explanations about priors and their comparison to the classical LASSO would help the reader.

AR: As requested, we have added this information. These sections now read:

> One limitation of the LASSO prior is that it introduces substantial bias in non-zero regression coefficients. To overcome this limitation, regularizing priors with better shrinkage properties have been developed. These priors still pull small regression coefficients towards zero, but exert less bias on larger regression coefficients. One example is the horseshoe prior (Carvalho, Polson, & Scott, 2010). It has heavier tails than the LASSO prior, which means that it does not shrink (and therefore bias) substantial coefficients as much. One remaining limitation of the horseshoe prior is that it is difficult to specify a prior that ensures a sufficiently sparse solution. The regularizing horseshoe was introduced to overcome this limitation; it uses information about the expected number of parameters to determine an appropriate shrinkage prior (Piironen & Vehtari, 2017).

RC: Why details relevant for the Bayesian estimation are included in the "Implementation" section? I mean lasso prior and horseshoe prior, that is, details useful to understand the methodology.

AR: We agree with the Reviewer that this information would be better positioned in another section. To address this comment, we have restructured the paper; the information about priors is now incorporated into the *Bayesian estimation* section, and the *Implementation* section now exclusively relates to the `pema` R-package.

RC: Can you comment about the choice of the numerical values in the priors at the end of page 11?

AR: Work in progress

RC: line 221: "this extension": which one?

AR: We intended to explain that the regularizing horseshoe prior is an extension of the (normal) horseshoe prior. This was insufficiently clear. To address this comment, we have written the following section:

> An alternative to the use of a shrinkage penalty is Bayesian estimation with a regularizing prior. Bayesian estimation combines information from the data with a *prior distribution*. The prior distribution assigns a-priori probability to different parameter values. Likely parameter values have a high probability density, and unlikely parameter values have a low probability density. The aforementioned (frequentist) approaches, by contrast, treat every possible parameter value as equally plausible. The prior distribution is updated with the likelihood of the data to form a posterior distribution, which reflects expectations about likely parameter values after having seen the data (for an extensive introduction, see **mcelreathStatisticalRethinkingBayesian2020?**).
>
> A regularizing prior distribution shrinks small coefficients towards zero by assigning high probability mass to near-zero values. There are many different regularizing prior distributions (Erp, Oberski, & Mulder, 2019). Some of these regularizing priors are analogous to specific frequentist methods. For example, a double exponential prior (hereafter: LASSO prior) results in posterior distributions whose modes are identical to the estimates from LASSO-penalized regression (Park & Casella, 2008). There are some notable differences between the frequentist LASSO penalty and its Bayesian counterpart, the LASSO prior. First, the frequentist LASSO penalty has a tuning parameter λ, whose value is usually chosen via cross-validation. In the Bayesian approach, an analogous role is fulfilled by a hyperparameter of the shrinkage prior. Instead of using cross-validation, however, it is possible to specify a hyperprior for this parameter, so that its value can be estimated from the data.
>
> One limitation of the LASSO prior is that it introduces substantial bias in non-zero regression coefficients. To overcome this limitation, regularizing priors with better shrinkage properties have been developed. These priors still pull small regression coefficients towards zero, but exert less bias on larger regression coefficients. One example is the horseshoe prior (Carvalho et al., 2010). It has heavier tails than the LASSO prior, which means that it does not shrink (and therefore bias) substantial coefficients as much. One remaining limitation of the horseshoe prior is that it is difficult to specify a prior that ensures a sufficiently sparse solution. The regularizing horseshoe was introduced to overcome this limitation; it uses information about the expected number of parameters to determine an appropriate shrinkage prior (Piironen & Vehtari, 2017).

RC: line 225: "This is accomplished by.." Why?

AR: We have rephrased this so that it is clearer that the formula explains how this is achieved. The section now reads:

> An attractive property of this shrinkage prior is that it can incorporate prior information regarding the expected number of relevant moderators. This is accomplished by having the scale of the global shrinkage parameter λ_0^2 depend on the expected number of relevant moderators p_0. The shrinkage parameter is then given by $\lambda_0^2 = \frac{p_0}{p-p_0}\frac{\sigma}{\sqrt{n}}$, where p represents the total number of moderators, σ is the residual standard deviation and n equals the number of observations.

RC: line 240: "values are reasonable in most applications" Why? Can you prove this?

AR: We thank the Reviewer for pointing out that this statement raises more questions than it answers. To address this Reviewer comment, we have moved this statement to the Discussion and slightly rephrased it to clarify

that it is supported by the results of the present study. It now reads:

> All simulations were conducted with default settings for the model's prior distributions. Our results suggest that these defaults are suitable for a wide range of situations, including when model assumptions are violated. However, bear in mind that model parameters are influenced by the prior distribution. It is good practice to perform sensitivity analysis to determine how sensitive the model results and inferences are to different prior specifications. Performing sensitivity analyses is particularly important when the sample is small, as in this case, the prior is more influential.

RC: lines 244-251 seems to be pertinent to the methodology of the proposal not to implementation.

AR: We agree with the Reviewer that the writing in this section was too methodological. To address this comment, we have made two changes: This paragraph has been moved to the section on Bayesian estimation, and we have rewritten it as follows:

> Unlike the frequentist LASSO algorithm, Bayesian regularized estimation does not shrink coefficients to be exactly equal to zero. Therefore, variables must be selected post-estimation. One way to do so is by the use of probability intervals. A moderator is selected when a 95% interval excludes zero. Two commonly used probability intervals in Bayesian inference are the credible interval and the highest posterior density interval (**mcelreathStatisticalRethinkingBayesian2020?**). The credible interval (CI) is the Bayesian counterpart of a confidence interval, and it is obtained by taking the 2.5% and 97.5% quantiles of the posterior distribution. The highest posterior density interval (HDPI) is the narrowest possible interval that contains 95% of the probability mass (**kruschkeChapter12Bayesian2015?**). When the posterior distribution is symmetrical, the CI and HDPI are the same. However, when the posterior is skewed, the HPDI has the advantage that all parameter values within the interval have a higher posterior probability density than values outside the interval. This suggests that the HPDI might be superior for performing inference on residual heterogeneity parameters, which have a skewed posterior distribution by definition. For inference on regression coefficients, the choice of interval is likely less crucial.

RC: More in general, I do not like to see a mixture of methodology and R code. I think I would be more useful for the reader to have sections describing the proposed approach and the underlying methodology (something that is not present at the emano) and then having an appendix about the code.

AR: To address this comment, we have made a stricter separation between methodology and code. All references to code are now relegated to the Implementation section.

We do believe that there are advantages to having code in the paper and not in an appendix, as the intended readership for this paper consists of those who wish to conduct a Bayesian regularized meta-analysis. We hope that the current approach is an acceptable compromise.

RC: line 278: "x==0" could be substituted by "x is equal to 0" as this is a text, not a source code.

AR: We have made the suggested change.

RC: Intercept. I don't think it is so relevant to deserve a section. Why not inserting a line in the section devoted to the code?

AR: This section is important because omitting the intercept from the model has very different implications for

penalized Bayesian meta-regression than for frequentist models (where the model is equivalent, regardless of whether the intercept is included or not).

To clarify this, we have rewritten this section to focus on the implications of including or excluding the intercept. The line explaining *how* to include or exclude the intercept has been moved to the Implementation paragraph, consistent with the Reviewer's request to relegate code to a separate section:

```
get_revision("intercept")
```

RC: Performance indicators, Design factors, Predictive performance * You split the data in training set and test set, for each iteration. So, how do you account for the variability associated to the choice of the test set (as done for example in cross-validation)?

AR: We respectfully believe that this comment is based on a misunderstanding of the procedure. We do not split a finite sample into a training and testing set - we simulate data for both sets from the same known data generating model. Each sample from this model is independent and identically distributed (IID). The distinction between training and testing is thus arbitrary. This is different from real data analyses, where the available dataset is finite, and training and testing data are drawn from that finite dataset - thus violating the assumption of IID.

In a real dataset, if one case has very unusual values (i.e., an outlier), the model could be biased if this case ends up in the training data, and the performance metrics could be biased if this case ends up in the testing data. In our simulation study, outliers have exactly equal probabilities of appearing in the training or testing data - and as we are not sampling from a finite dataset, the probability of having an outlier in the training data is independent from having one in the testing data.

Their effect on the estimated parameters is known as "Monte Carlo error" and averages out across replications of the simulation.

RC: I'm surprised by the large number fo linea devoted to the explanation of such a basic indicator like R2. In addition, it is well known to be optimistic, and thus the adjusted version I usually preferable. Why don't you evaluate the method using other indicators, such as the adjusted r2, AIC, or BIC?

AR: We respectfully feel that it is important to define how we operationalized each outcome metric. In this case, the metric of interest is *predictive* R^2, that is to say, variance explained in a new sample. This metric helps assess whether the algorithms overfit the data - which is the concern we try to address with our method. For example, we see the RMA analysis overfit, especially when the true effect size is zero. Apparently, despite the many lines devoted to explaining our outcome metric, this was still not sufficiently clear.

The reason we do not use AIC or BIC is because these are comparative fit indices, only suitable for comparing a set of models; we have only a single model and AIC or BIC do not speak to the absolute fit of a single model.

The reason we do not use adjusted R^2 is because it is an estimate of the predictive R^2 based on only a training sample. As we simulate data from a known true model, we do not need to *estimate* predictive R^2 - we can *calculate* it directly, based on a new sample.

Work in progress

RC: line 344: I suppose you mean the estimate of sensitivity and specificity.

AR: We have rephrased this as suggested.

RC: **line 368: 100 datasets only? Probably, there is a typo, as 100 is the number of datasets included in the test set, if I'm not wrong. In any case, 100 is a very low number of replicates for a simulation study, I suggest to increase at least to the common value equal to 1,000.**

AR: As these simulations are very computationally demanding, and increasing the number of replications adds only marginal accuracy gains, we did indeed use 100 replications. However, as per the Reviewer's request, we have increased this to 1000.

RC: **line 385: do you mean "Table" 1?**

AR: Thank you for pointing out this omission; we have corrected it.

RC: **line 405: "the" most negative?**

AR: We corrected this, and slightly rephrased the sentence for clarity.

RC: **lines 449-452 are not necessary as bias and variance are known concepts. However, in Tables I can see only bias, and not information about variance.**

AR: We respectfully feel that it is prudent to define how all outcome metrics are operationalized. Even though the general concepts of bias and variance may be known to most readers, we cannot assume that they will know exactly how these are calculated in the context of this simulation study - particularly for the broad readership of a journal like Research Synthesis Methods.

RC: **lines 469-470: there is a white space**

AR: We thank the Reviewer for pointing this out; a fragment of a sentance was cut here. We restored the original text.

RC: **Could you please avoid writing pema::bonapersona and use instead dataset titled bonapersona in the R pema package? This is not a list of lines code, but a text.**

AR: We have done as requested.

RC: **you have 440 experiments in the dataset: do you mean 440 study included in the meta-regression? If so, the number is much. Larger than the scenarios evaluated in the simulations.**

AR: This observation is correct, but this is not a problem. The method is validated for up to 100 studies, so it should also be valid for >100 studies, as a larger sample size gives more power.

The reason we used these data is because we were able to secure permission from the original author to include them in the R-package. This means we can make the examples reproducible, requiring only the `pema` package. We feel that it adds a lot to the paper to present a reproducible real-data example based on a novel dataset. To address this comment, we now acknowledge in the text that this sample is relatively large compared to the sample sizes of the simulation study:

Work in progress

RC: **Finally, There are many typos and inconsistencies throughout the text (e.g., data set and dataset, random effect and randoms-effects, . . .).**

AR: **Work in progress**

2. Reviewer 2

RC: **Comments to the Author The authors investigated selection of moderators for the meta-regression, which is especially useful to explain the sources of heterogeneity between trials. They proposed the use of regularizing priors for meta-regression within a Bayesian framework. They shared publicly available R package "pema" which relies on rstan, which helps practitioners to use the proposed methods. I think the contribution of the paper and R package is important. However, I have some comments.**

AR: We thank the Reviewer for their encouraging words and thought-provoking comments, which we feel have helped improve the quality of the manuscript.

[RC 2.1.] Major points

RC: **1) I think it is important to note some limitations of meta-regression. For example, the use of moderators might "break" the randomization, when the studies analyzed are randomized controlled trials. This is because it is not possible to randomize patients to one moderator, see Thompson and Higgins (2002) section 3 for further discussion of limitations of meta-regression of clinical trials.**

AR: **Work in progress**

RC: **2) Important references are missing, in which the use of regularization or penalization approaches for the meta-analysis were investigated, although not the same purpose as in the current paper. For example, in Chung et al (2013) regularizing prior was used for the heterogeneity parameter to avoid boundary estimates, and in Günhan et al (2020) regularizing priors were used for the heterogeneity parameter and treatment effect parameter to deal with data sparsity. Finally and most importantly, Röver et al (2021) reviewed different use of regularizing priors and provided some guidance. Note that in the mentioned references, the term of weakly informative priors are used instead of regularizing priors (also see an earlier reference by Gelman (2006) in a hierarchical model context). Mention of these publications and relationship to the present paper can help the reader.**

AR: We thank the Reviewer for suggesting these references, which had eluded our literature search.

We agree that the use of weakly informative priors for heterogeneity parameters was insufficiently discussed in the previous version of our manuscript, and have now added a paragraph on that topic (see our response to the next comment).

However, it is important to note that the use of weakly informative priors (WIP) for heterogeneity parameters serves a categorically different purpose than the use of regularizing priors for regression coefficients - even if the term “regularization” has been used for both in prior literature. The purpose of WIP for heterogeneity parameters is to aid model convergence and avoid boundary estimates. The purpose of regularizing priors for regression coefficients is to perform variable selection. We believe that the newly added paragraph sufficiently clarifies these distinct uses of “regularization” (see response to next comment).

Thank you for pointing out these relevant references. We have included the reference to Chung et al. (2013), Gelman (2006) and Rover et al. (2021) in the new section on the prior for τ (see the next comment). In addition we have added the following section on page ... to be more explicit about the terminology of different types of priors:

Suggest to add this paragraph in the section on Bayesian estimation after “which reflects expectations about likely parameter values after having seen the data.” Note: if included, this should be updated in the text from a

previous comment to reviewer 1 as well. Also here it is relevant whether we use "regularizing" or "shrinkage" priors and should mention the definite choice explicitly.

"As a result, the posterior distribution is a compromise between the likelihood of the data and the prior distribution. The influence of either the data or the prior is determined by the sample size as well as the informativeness of the prior distribution. The larger the sample size, the more influential the likelihood of the data will be. Similarly, as the prior distribution becomes more informative, i.e., more peaked, it will exert more influence on the results. Priors can thus be placed on a continuum from"non-informative" to "highly informative". Weakly informative or regularizing priors are often used in situations when there is some prior information available, for example regarding the boundaries of the parameter space or, as is the case here, regarding the existence of parameters equal to zero. Throughout this manuscript, we will use the term "shrinkage prior" to refer to the weakly informative or regularizing prior that holds the prior knowledge that some coefficients are equal to zero and should therefore be shrunken towards zero."

Work in progress

RC: 3) An important part of Bayesian random-effects meta-analysis is the specification of the prior for the heterogeneity parameter tau. Can you include more specifications on the prior choice for tau and influence of the choice to the results (if there is any).

AR: We agree with the Reviewer that the prior specification for the heterogeneity parameter tau is important and should be addressed in the text. To address this comment, we have included the following paragraph in the revision:

The choice of prior distributions is an important decision in any Bayesian analysis. This also applies to the heterogeneity parameters. In the case of random effects meta-regression, the only heterogeneity parameter is the between-studies variance, τ^2. In the case of three-level multilevel meta-regression, there is a within-study and between-studies variance.

A crucial challenge with heterogeneity parameters in meta-regression is that the number of observations at the within- and between-study level is often small. This can result in poor model convergence (**roverWeaklyInformativePrior2021?**), or boundary estimates at zero (**chungAvoidingZeroBetweenstudy2013?**). A well-known advantage of Bayesian meta-analysis is that it can overcome these challenges by using weakly informative priors, which guide the estimator towards plausible values for the heterogeneity parameters. There is less consensus, however, about which priors are best for this purpose (**roverWeaklyInformativePrior2021?**). BRMA uses a prior that was specifically developed for multilevel heterogeneity parameters (**gelmanPriorDistributionsVariance2006?**): the half-Student's t distribution with large variance, $df = 3, scale = 2.5$. Note that other suitable weakly informative priors have been discussed in the literature, such as the Wishart prior (**chungWeaklyInformativePrior2015?**). There has also been increasing interest in the use of informative priors for heterogeneity parameters (**thompsonGroupspecificPriorDistribution2020?**). Informative priors incorporate substantive domain knowledge about likely values for the parameter. Informative priors are different from weakly informative priors, which merely guide the estimator towards possible values (e.g., by excluding negative values for the variance), or restrict it to plausible values to aid model convergence. BRMA takes a pragmatic approach to Bayesian analysis, using weakly informative priors to aid convergence for heterogeneity parameters, and regularizing priors to perform variable selection for regression coefficients. The use of informative priors is out of scope for BRMA. If researchers do wish to construct alternative prior specifications, they may want to develop a custom model in `rstan` instead (Stan Development Team, 2022).

Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med. 2013;32(23):4071-4089.

Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 2006;1(3):515-534. https://doi.org/10.1214/06-BA117A.

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E. J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123-138.

Röver, C, Bender, R, Dias, S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Syn Meth. 2021; 12: 448– 474. https://doi.org/10.1002/jrsm.1475

RC: 4) In the abstract and in the main text, it states "We present a simulation study to validate the performance of BRMA relative to state-of-the-art meta-regression (RMA)". What does RMA refer to here, is it "Random effect Meta-Analysis using REML"? If yes, I think the term random-effect meta-analysis (or meta-regression) using RMLE is more clear than state-of-the-art meta-analysis (or meta-regression).

AR: We have made the suggested correction. Both in the Abstract and in the text, the first time we introduce RMA, we introduce it as "random effects meta-analysis using restricted maximum likelihood".

[RC 2.2.] Minor points

RC: Line 385: "see 1": I think Table is missing

AR: Thank you; we have corrected this.

3. References

1) Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted?. Stat Med. 2002;21:1559-1573.
2) Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med. 2013;32(23):4071-4089.
3) Günhan, BK, Röver, C, Friede, T. Random-effects meta-analysis of few studies involving rare events. Res Syn Meth. 2020; 11: 74– 90. https://doi.org/10.1002/jrsm.1370
4) Röver, C, Bender, R, Dias, S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Syn Meth. 2021; 12: 448– 474. https://doi.org/10.1002/jrsm.1475
5) Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 2006;1(3):515-534. https://doi.org/10.1214/06-BA117A.

4. References

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Erp, S. van, Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Jargowsky, P. A. (2004). The Ecological Fallacy. In K. Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement* (Vol. 1, pp. 715–722). San Diego, CA: Academic Press.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-ejs1337si

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437–437. https://doi.org/10.1038/470437a

Stan Development Team. (2022). *RStan: The R interface to Stan*. Retrieved from https://mc-stan.org/