Selecting relevant moderators with Bayesian regularized meta-regression

Caspar J. Van Lissa[1], Sara van Erp[2], & Eli-Boaz Clapper[2]

[1] Tilburg University, dept. Methodology & Statistics

[2] Utrecht University, dept. Methodology & Statistics

Author Note

This is a preprint paper, generated from Git Commit # 841d1dd. This work was funded by a NWO Veni Grant (NWO Grant Number VI.Veni.191G.090), awarded to the lead author.

The authors made the following contributions. Caspar J. Van Lissa: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing; Sara van Erp: Methodology, Software, Writing – original draft, Writing – review & editing; Eli-Boaz Clapper: Formal Analysis, Writing – original draft, Writing – review & editing.

Correspondence concerning this article should be addressed to Caspar J. Van Lissa, Professor Cobbenhagenlaan 125, 5037 DB Tilburg, The Netherlands. E-mail: c.j.vanlissa@tilburguniversity.edu

Abstract

When meta-analyzing heterogeneous bodies of literature, meta-regression can be used to account for potentially relevant between-studies differences. A key challenge is that the number of candidate moderators is often high relative to the number of studies. This introduces risks of overfitting, spurious results, and model non-convergence. To overcome these challenges, 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. This method is suitable when there are many potential moderators, but it is not known beforehand which of them are relevant. A simulation study compared BRMA against state-of-the-art random effects meta-regression using restricted maximum likelihood (RMA). Results indicated that BRMA outperformed RMA on three metrics: BRMA had superior predictive performance, which means that the results generalized better; BRMA was better at rejecting irrelevant moderators, and worse at detecting true effects of relevant moderators, while the overall proportion of Type I and Type II errors was equivalent to RMA. BRMA regression coefficients were slightly biased towards zero (by design), but its residual heterogeneity estimates were less biased than those of RMA. BRMA performed well with as few as 20 studies, suggesting its suitability as a small sample solution. We present free open source software implementations in the R-package `pema` (for penalized meta-analysis) and in the stand-alone statistical program JASP. An applied example demonstrates the use of the R-package.

Keywords: meta-analysis, machine learning, regularization, bayesian, lasso, horseshoe

Word count: 5356

Selecting relevant moderators with Bayesian regularized meta-regression

A common application of meta-analysis is to summarize existing bodies of literature. A crucial challenge is that there is often substantial heterogeneity between studies, because similar research questions are studied in different labs, sampling from different populations, and using different study designs, instruments, and methods. Any of those between-studies differences can introduce *systematic heterogeneity* in observed effect sizes. Suspected causes of systematic heterogeneity can either be used as exclusion criteria, or controlled for using *meta-regression* (see López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014). The latter approach provides an opportunity to learn which factors impact the effect size found. However, a limitation of meta-regression is that it requires a relatively high number of cases (studies) per parameter to obtain sufficient statistical power. In applied meta-analyses, the number of available studies is often low (Riley, Higgins, & Deeks, 2011). This introduces a risk of overfitting, which results in uninterpretable model parameters (Hastie, Tibshirani, & Friedman, 2009). In extreme cases, the ratio of cases to parameters can be so low that the model is not (empirically) identified, resulting in non-convergence (Akaike, 1974). Accounting for between-studies heterogeneity thus poses a non-trivial challenge to classic meta-analytic methods. The risk of arriving at false-positive conclusions when there are many potential moderators is so ubiquitous that it was referred to as the "primary pitfall" in meta-regression (S. G. Thompson & Higgins, 2002). The present study introduces a novel method to overcome this pitfall by imposing Bayesian regularizing (LASSO and regularized horseshoe) priors on the regression coefficients. These priors shrink the effect of irrelevant predictors towards zero while leaving important predictors relatively unaffected. The result is a sparse model with fewer non-zero parameters, which benefits model convergence, reduces overfitting, and helps identify relevant between-study differences.

Variable selection

The "curse of dimensionality" refers to the aforementioned problems that arise when the number of moderators is high relative to the number of cases (Hastie et al., 2009). It can be overcome by performing *variable selection*: identifying a smaller subset of relevant moderators from the larger set of candidate moderators (Hastie et al., 2009). Prior authors have stressed the need to perform variable selection in meta-regression, for example, by limiting the number of moderators considered (S. G. Thompson & Higgins, 2002). This does not resolve the problem, however, as failing to consider a moderator does not mean that it is irrelevant. Instead, moderators ought to be selected based on their theoretical or empirical relevance for the studied effect.

One approach is to select variables on theoretical grounds. 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 are well-known (Baker et al., 2009; S. G. Thompson & Higgins, 2002): For example, meta-regression may find a significant positive effect of average sample age on the effect size of a randomized controlled trial, even if age is uncorrelated with treatment efficacy within each study. Less well-known is that the same problem applies when using individual level theory to select study level moderators: If theory states that an individual's age influences their susceptibility to treatment, that does not imply that average sample age will be a relevant moderator of study-level treatment effect in meta-regression. One rare example of study level theory is the *decline effect*: effect sizes in any 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 then 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. At present, few such study level theories about the drivers of heterogeneity exist, and until they are developed, theory has limited utility for variable selection. A further complication is that theoretically relevant variables might not be reported in many published papers, which may be one reason why moderator analyses are rarely executed as planned (S. G. Thompson & Higgins, 2002)

An alternative solution is data-driven variable selection using appropriate statistical techniques. This is a focal issue in the discipline of machine learning (Hastie et al., 2009). There is precedent for the use of machine learning for variable selection in meta-analysis (Van Lissa, 2020). This prior work used the non-parametric *random forest* algorithm. A limitation of random forests is that its results are harder to interpret than linear models, which describe the effect of each moderator with a single parameter. The present study instead uses *regularization*, which is a method for variable selection in linear models. Regularization shrinks small model parameters towards zero, such that irrelevant moderators are eliminated. Different approaches to regularization exist (Hastie et al., 2009). 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 regularized horseshoe prior.

Statistical underpinnings

To understand how BRMA estimates the relevant parameters and performs variable selection, it is instructional to first review the statistical underpinnings of classical meta-analysis. In its simplest form, meta-analysis amounts to computing a weighted average of the effect sizes. Each effect size is assigned a weight that determines how influential it is in calculating the summary effect. The weights are based on specific assumptions. For example, the *fixed effect* model assumes that each observed effect size T_i is an estimate of an underlying true population effect size β_0 (Hedges & Vevea, 1998). This assumption is appropriate when meta-analyzing close replication studies (Higgins, Thompson, & Spiegelhalter, 2009). The only cause of heterogeneity in observed effect sizes is presumed to be sampling error, v_i, which is treated as known, and computed as the square of the standard error of the effect size. Thus, for a collection of k studies, the observed effects sizes of individual studies i (for $i \in [1, 2, \ldots k]$) are given by:

$$T_i = \beta_0 + \epsilon_i \tag{1}$$

$$\text{where } \epsilon_i \sim N(0, v_i) \tag{2}$$

The estimated population effect size $\hat{\beta}_0$ is then a weighted average of the observed effect sizes. The assumption that sampling error is the only source of variance in observed effect sizes implies that studies with smaller standard errors estimate the underlying true effect size more precisely and should accrue more weight. Therefore, fixed effect weights are simply the reciprocal of the sampling variance, $w_i = \frac{1}{v_i}$. The estimate of the true effect is a weighted average across observed effect sizes:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{3}$$

The *random effects* model, by contrast, assumes that, in addition to sampling error, true effects vary for random reasons, and thus follow a normal distribution with mean β_0 and variance τ^2 (Hedges & Vevea, 1998). This assumption is appropriate when studies are conceptually similar but differ in small random ways (Higgins et al., 2009). The observed effect sizes are thus given by:

$$T_i = \beta_0 + \zeta_i + \epsilon_i \tag{4}$$

$$\text{where } \zeta_i \sim N(0, \tau^2) \tag{5}$$

$$\text{and } \epsilon_i \sim N(0, v_i) \tag{6}$$

In this model, the error term ζ_i represents between-studies heterogeneity, with variance τ^2. As in the fixed effect model, studies with smaller sampling errors are assigned more weight. In contrast to the fixed effect model, the random effects model assumes that all studies provide some information about the underlying distribution of true effect sizes. Fixed effect weights would discount the information smaller studies provide about the scale of this distribution, which is represented by its variance τ^2. To overcome this limitation, the weights are attenuated in proportion to the variance. The random effects weights are thus given by $w_i = \frac{1}{v_i + \hat{\tau}^2}$. Whereas sampling error is still treated as known, the between-study heterogeneity τ^2 must be estimated. This estimate is represented by $\hat{\tau}^2$.

Between-studies heterogeneity is not always random, however. Meta-regression extends the random effects model to account for systematic sources of heterogeneity, which are coded as moderators. It estimates the effect of moderators on effect size, and provides an estimate of the overall effect size and residual heterogeneity after controlling for their influence. For example, if studies have been conducted in Europe and the Americas, one could code a binary moderator variable called "continent". Using meta-regression, one can then estimate either the continent-specific effect size, or control for the difference between continents when estimating the overall average effect size. Similarly, if studies have examined the effect of a drug at different dosages, one could code dosage as a continuous moderator and estimate the overall effect size at average dosages, or at a specific dosage. The equation below describes a general model for p moderators, where $x_{1 \ldots p}$ represent the moderator variables, and $\beta_{1 \ldots p}$ their regression coefficients. Note that β_0 now represents the intercept of the distribution of true effect sizes after controlling for the moderators. This is a mixed-effects model; the intercept and effects of moderators are treated as fixed and the residual heterogeneity as random (López-López et al., 2014):

$$T_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \zeta_i + \epsilon_i \tag{7}$$

Regularized regression. Meta-regression models are most commonly estimated using weighted least squares, with weights determined according to the aforementioned random effects model, where residual heterogeneity is estimated using restricted maximum likelihood (Patterson & Thompson, 1971; Viechtbauer, 2005). This approach, henceforth referred to as RMA, has low bias, which means that, across hypothetical replications of a study, the average values of model parameters are close to their true population values (Panityakul, Bumrungsup, & Knapp, 2013). Low bias comes at the cost of higher variance, however, which means that the estimated values of population parameters vary more from one hypothetical replication to the next. This phenomenon is known as the *bias-variance trade-off*. In general, an estimator with low bias and high variance produces results that generalize less well to new data than an estimator with high bias and low variance. Regularized approaches to regression intentionally increase bias in order to reduce variance. This is a sensible approach in the context of small samples, which are common in meta-analyses, because small samples incur a high risk of overfitting, and typically have relatively high levels of multicollinearity, due to the higher probability that extreme values on one moderator coincide with extreme values on another. In such cases, regularized regression reduces the risk of overfitting and increase generalizability of the results (see Hastie et al., 2009).

To understand how regularized regression introduces bias, consider a comparison between ordinary least squares regression and LASSO regression (for a more elaborate introduction, see Tibshirani, 2011). Ordinary least squares regression (OLS) estimates model parameters by minimizing the Residual Sum of Squares (RSS) of the outcome variable, given by the formula below. In this equation, n is the number of participants; y_i is the outcome variable, and x is one of p predictor variables.

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{8}$$

The resulting parameter estimates perfectly describe linear relations in the present data set, but generalize less well to new data. Regularized regression biases parameter estimates towards zero by adding a penalty term to the RSS. Most common is the LASSO penalty, which consists of the sum of the absolute regression coefficients, or the L1 norm (Hastie et al., 2009). As the LASSO penalty is a function of the regression coefficients, it increases when they get bigger. This incentivizes the optimizer to keep the regression coefficients as small as possible. The amount of regularization can be controlled by multiplying the penalty by a tuning parameter, λ. If λ is zero, the shrinkage penalty has no impact. As λ increases, all coefficients shrink towards zero, ultimately producing the null model. Cross-validation is often used to find the optimal value for the penalty parameter λ. The LASSO penalized residual sum of squares is given by:

$$PRSS = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

Note that many other regularizing penalties exist. This introduction focuses on the LASSO penalty because it is most ubiquitous, easy to understand, and has an analogue in Bayesian estimation, as explained in the next section.

Some seminal studies have applied the LASSO to perform moderator selection in meta-regression (Requia et al., 2018; Rosettie et al., 2021; Sebri & Dachraoui, 2021). This suggests that others have recognized its potential for exploring heterogeneity when the number of moderators is relatively high to the number of studies. However, these existing publications have taken a two-step approach, whereby moderators are first selected using LASSO regression, and selected moderators are then included in meta-regression analysis. This approach is fraught; firstly, because of known problems of inference after variable selection (Zhang, 1992). As the moderators included in the second step are based on an exploratory first step, their parameters are not valid for inference. Secondly, although the LASSO model in the first step accounts for potential multicollinearity by including all collinear variables but restricting the size of their coefficients, the meta-regression in the second step no longer does so. A three-step extension of the two-step approach exists that uses principles from the causal inference literature to overcome these limitations (Belloni, Chernozhukov, & Hansen, 2014). BRMA, by contrast, overcomes these limitations by introducing a one-step approach that performs inference within the penalized framework.

Bayesian estimation. An alternative to the use of a shrinkage penalty is Bayesian estimation with a regularizing prior. Whereas the aforementioned (frequentist) approaches treat every possible parameter value as equally plausible, Bayesian estimation combines information from the data with a *prior distribution* that 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 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 a more extensive introduction to Bayesian estimation, see McElreath (2020).

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, some of which are analogous to specific frequentist methods (van Erp, Oberski, & Mulder, 2019). 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). Both the frequentist LASSO penalty and the Bayesian LASSO prior have a tuning parameter λ that controls the amount of regularization. In frequentist LASSO, its value is usually chosen via cross-validation (Hastie et al., 2009). In the Bayesian approach, by contrast, a diffuse hyperprior can be used to optimize its value during model estimation (Park & Casella, 2008).

One limitation of the LASSO prior is that it biases all regression coefficients towards zero - for relevant as well as irrelevant moderators. 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. Two limitations of the horseshoe prior are 1) it lacks a formal way to include prior information regarding the degree of sparsity; and 2) it does not regularize coefficients far from zero. While the second problem is often considered a strength of the horseshoe prior, it can result in convergence issues when parameters are weakly identified. The regularized horseshoe was introduced to overcome these limitations (Piironen & Vehtari, 2017b).

The BRMA method introduced here offers both LASSO and regularized horseshoe priors. The LASSO prior is given by:

$$\beta_j \sim \text{DE}(0, \frac{s}{\lambda}) \tag{10}$$

$$\frac{1}{\lambda} \sim \chi^2(0, \nu_1) \tag{11}$$

where DE denotes the double exponential distribution centered around zero, with a scale determined by a global scale parameter s that is multiplied by the inverse of tuning parameter λ. Increasing the scale parameter extends the prior to cover more extreme values. The inverse tuning parameter is estimated from the data by assigning it a diffuse hyperprior: a χ^2 prior distribution with mean zero and degrees of freedom ν_1 (Park & Casella, 2008). Increasing the degrees of freedom assigns greater probability mass to extreme values, resulting in less regularization. The present study used default values for the prior parameters, $s = 1$, $\nu_1 = 1$, were chosen as a diffuse hyperprior (Park & Casella, 2008).

The regularized horseshoe prior combines global and local shrinkage of the regression coefficients with a finite slab that curtails the occurrence of very extreme values (Piironen & Vehtari, 2017b). For regression coefficients β_j, for $j \in [1 \ldots p]$ where p is the total number of moderators, the regularized horseshoe prior is given by:

$$\beta_j \sim N(0, \tilde{\tau}_j^2 \lambda), \text{ with} \tag{12}$$

$$\tilde{\tau}_j^2 = \frac{c^2 \tau_j^2}{c^2 + \lambda^2 \tau_j^2} \tag{13}$$

$$\lambda \sim \text{student-}t^+(\nu_1, 0, \lambda_0^2) \tag{14}$$

$$\tau_j \sim \text{student-}t^+(\nu_2, 0, 1) \tag{15}$$

$$c^2 \sim \Gamma^{-1}(\frac{\nu_3}{2}, \frac{\nu_3 s^2}{2}) \tag{16}$$

Note that global shrinkage parameters, which are not subscripted, affect all regression coefficients. Local parameters are indicated by subscript $_j$, and affect each individual regression coefficient separately. In these equations, N denotes the normal distribution, student-t^+ denotes the positive half of a t distribution, and Γ^{-1} denotes the inverse Gamma distribution. In this formula, λ controls the overall scale of the priors for all regression coefficients, where larger values for the global scale parameter λ_0^2 widen the range of values covered by the priors. The global degrees of freedom ν_1 control the overall thickness of the tails, with higher values resulting in thinner tails, which assigning less probability mass to extreme values. Lighter tails can aid model convergence when the model is weakly identified; for example, when there are more moderators than observations. The prior τ_j controls the local shrinkage of specific regression coefficients; its scale is fixed, but its degrees of freedom ν_2 control the incidence of extreme values in a similar way as ν_1. A finite "slab" applies additional regularization to very large coefficients, which provides greater numerical stability of the model. This slab is governed by a degrees of freedom parameter ν_3 and a scale parameter s. As before, increasing ν_3 assigns less probability mass to extreme values. Increasing s increases the range of values covered by the slab.

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 calculating the scale of the global shrinkage parameter λ_0^2 based on the expected number of relevant moderators p_{rel}. The shrinkage parameter is then given by $\lambda_0^2 = \frac{p_{rel}}{p - p_{rel}} \frac{\sigma}{\sqrt{n}}$, where σ

is the residual standard deviation and n equals the number of observations. The present study used default values for the prior parameters, as proposed by its authors: $\lambda_0^2 = 1$, $\nu_1 = 1$, $\nu_2 = 1$, $\nu_3 = 4$, and $s = 2$ (Piironen & Vehtari, 2017b).

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 (Röver et al., 2021), or boundary estimates at zero (Chung, Rabe-Hesketh, & Choi, 2013). 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 most suitable for this purpose (Röver et al., 2021). BRMA uses a prior specifically developed for multilevel heterogeneity parameters (Gelman, 2006): a half-Student's t distribution with large variance, student-$t^+(3, 0, 2.5)$. Note that other relevant weakly informative priors have been discussed in the literature, such as the Wishart prior (Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2015). There has also been increasing interest in the use of informative priors for heterogeneity parameters, which incorporate substantive knowledge about plausible parameter values (C. G. Thompson & Becker, 2020). Informative priors exert substantial influence on the parameter estimates. They thus differ from weakly informative priors, which restrict the estimator towards possible values (e.g., by excluding negative values for the variance), or guide it towards 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).

The frequentist LASSO algorithm shrinks coefficients to be exactly equal to zero, and thus inherently performs variable selection. Other approaches to regularization - frequentist or Bayesian - lack this property. However, an advantage of the Bayesian approach is that its posterior distributions lend itself to exact inference. One can use probability intervals to determine which population effects are likely non-zero; for example, by selecting moderators whose 95% interval excludes zero. Two commonly used Bayesian probability intervals are the credible interval and the highest posterior density interval (McElreath, 2020). 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 (Kruschke, 2015). 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.

Standardizing predictors. As explained in Formula (9), regularization penalizes all coefficients equally, without regard for their scale. If variables are on different scales, this can lead to uneven penalization of coefficients in which variables with smaller standard deviations are biased more strongly towards zero (Lee, 2015). If the scale of predictor x is increased by a factor 10, its regression coefficient is reduced by a factor 10, bringing it closer to zero where it will be more affected by penalization. Standardization is a widely used method for equalizing predictor scales (Gelman, 2008). Standardization is a linear transformation that sets the mean of all predictors to 0 and their standard deviation to 1. Like most other regularizing methods, BRMA performs standardization by default (Tibshirani, 1996). After parameters are estimated using standardized variables, they can be restored to their original scales. For the intercept, the transformation is:

$$b_0 = b_{0Z} - \mathbf{b}_Z \frac{\bar{\mathbf{x}}}{\mathbf{s}_X}$$

where b_0 is the intercept, b_{0Z} is the intercept for the standardized predictors, $\bar{\mathbf{x}}$ and $\mathbf{s}_x$ are the vectors of predictor means and variances, and $\mathbf{b}_\mathbf{Z}$ is the vector of regression coefficients for the standardized predictors. The regression coefficients are returned to their original scale by applying:

$$\mathbf{b}_x = \frac{\mathbf{b}_z}{\mathbf{s}_x}$$

Note that standardization is not always necessary or desirable. Standardization is not necessary if predictors are already on equivalent scales, in which case penalization already affects them all equally. There are additional considerations regarding standardization of categorical predictors (Alkharusi, 2012). As binary predictors can be straightforwardly included as predictors in linear models, the most common way to represent categorical predictors is by choosing one response option as reference category, and creating binary dummy variables to represent other response categories. If these dummies are not standardized, they might be unevenly penalized, as explained before. However, standardizing dummy variables compromises the interpretability of their regression coefficients (Tibshirani, 1997; Wissmann, Toutenburg, et al., 2007). To illustrate this challenge, consider bivariate regression with a single binary predictor x that takes on values 0 and 1 predicting outcome y. The intercept represents the expected value of y when x is equal to zero, and the regression coefficient represents the difference in the expected value of y between the two conditions (Alkharusi, 2012). By standardizing this binary predictor, the reference value is no longer zero, and both the intercept and its regression coefficient have no clear interpretation anymore. Extending this example to the multivariate case further complicates the problem (Wissmann et al., 2007). The appropriate solution depends on the research goals; if the primary goal is variable selection, then the dummies should be standardized. However, if the primary goal is interpretation of the coefficients, they should not be (Gelman, 2008). A related challenge is that, whereas various coding schemes for categorical predictors are equivalent in standard linear regression, in penalized regression, the coding scheme does affect model fit and interpretation of the coefficients (Chiquet, Grandvalet, & Rigaill, 2016; Detmer, Cebral, & Slawski, 2020).

Intercepts. The general linear model used in BRMA typically includes an intercept, which reflects the expected value of the outcome when all predictors are equal to zero, and regression coefficients for the effect of moderators. If the analysis contains categorical predictors, it may be desirable to omit this intercept. To understand why, first consider the model with an intercept. Standard practice is to encode category membership with dummy variables, with values $x \in \{0, 1\}$. For a variable with c categories, the number of dummy variables is equal to $c - 1$. The omitted category functions as a reference category, and its expected value is represented by the model intercept b_0. The regression coefficients of the dummy variables, $b_{1 \ldots c}$, indicate the difference between the expected values of the reference category and of the category represented by the dummy. This is useful when there is a meaningful reference category. For example, imagine a study on the effectiveness of interventions for specific phobia with two interventions: Treatment as usual, and a novel intervention. In this case, it makes sense to code treatment as usual as the reference category, and dummy-code the new contender. The intercept b_0 then represents the average effect size of treatment as usual, and the effect of the dummy b_1 indicates whether the newly developed intervention has a significantly different effect size from treatment as usual. In other cases, there may not be a straightforward reference category. For example, imagine a study on the effectiveness of one intervention for specific phobia in two continents. In this case, it makes more sense to estimate the average effect for all continents separately - in other words, to conduct a multi-group analysis. This is achieved by removing the intercept, and including all c dummy variables. In the context of standard linear regression, both approaches are equivalent, but in regularized regression, shrinkage affects the intercept differently from the dummy variables. Consequently, a reasoned choice must be made about whether to include an intercept or not.

Implementation

To facilitate adoption of the BRMA method in applied research, we have implemented it in two software packages. First, in the statistical programming language R (Team, 2022). R-users can install the package `pema`, short for *pe*nalized *m*eta *a*nalysis, from CRAN by running `install.packages("pema")`. Second, non-R-users can use BRMA via a graphical interface in the free, open source statistical program JASP (JASP Team, 2022) via the menu option “Penalized Meta-Analysis”, see Figure 1.

For estimation, `brma()` depends on Stan, a probabilistic programming language that uses Hamiltonian Monte Carlo to sample from the posterior distribution (Stan Development Team, 2022). Stan is written in C++, and thus computationally efficient, but custom models must be compiled prior to estimation. Installing a toolchain to compile models requires some technical sophistication, which potentially restricts the user base. Moreover, model compilation adds unnecessary computational overhead for standard applications. To overcome these limitations, the `pema` package includes pre-compiled stock models with opinionated default options. At the time of writing, these include random effects and three-level meta-regression with and without an intercept. R-users can refer to the package documentation to see what options are available at the time of reading by running `?pema::brma`. Researchers who wish to construct a model that is currently out of scope of `brma()` are referred to `rstan` instead (Stan Development Team, 2022). As a starting point, the `rstan` source code for the stock models included with `pema` can be accessed by running `pema:::stanmodels`. We welcome user contributions of additional models.

The function `brma()` has two main interfaces: a formula interface, corresponding to base-R functions like `lm()`, which allows the user to specify a model `formula` that references variables in a `data` argument. The second interface is more amenable to machine learning applications, and accepts an `x` matrix of predictors and a `y` vector of effect sizes. Additionally, `brma()` has an argument `vi`, which refers to the effect size variances, and `study`, which (optionally) refers to a clustering variable for three-level meta-regression. Both of these arguments accept either the name of a column in `data`, or a numeric vector.

As mentioned before, the R-implementation of BRMA has several options that can be customized. The most important option relates to the choice of priors for the regression coefficients. At the time of writing, `brma()` supports two priors for regression coefficients: the LASSO and the regularized horseshoe. A prior is selected using the `method` argument; the `prior` argument is used to specify custom values for the prior hyperparameters (see Statistical underpinnings). The parameters of the LASSO prior are explained in Equation (11), and those of the regularized horseshoe in Equation (13). Table 1 provides an overview of the arguments that can be passed to `prior` to control these parameters, along with a rudimentary description of the effect of increasing the value of each parameter.

Standardization is an important step in Bayesian regularized meta-analysis, as explained before. By default, `brma()` standardizes the predictor matrix, and restores model coefficients to their original scale, as explained in Statistical underpinnings. There are two ways to circumvent this default standardization. The first is to disable standardization entirely, analyzing predictors in their original scale, by setting `standardize = FALSE`. Alternatively, `brma()` allows custom standardization. To use this option, first manually standardize (some of) the predictors. Then, when calling `brma()`, provide the means (`means`) and standard deviations (`sds`) that should be used to restore coefficients to the predictors' original scale. This can be accomplished using the argument `standardize = list(center = means, scale = sds)`. For predictors that should not be standardized, simply pass a mean of 0 and a standard deviation of 1; this leaves the coefficient in question unaffected.

Simulation study

We performed a simulation study to validate the BRMA algorithm. As a benchmark for comparison, we used random effects meta-regression with restricted maximum likelihood estimation (RMA, Viechtbauer, 2010), which is the current state-of-the-art in the field. We evaluated the algorithms' predictive performance in new data, ability to perform variable selection, and ability to recover population parameters. Our research questions are whether BRMA offers a performance advantage over RMA in terms of any of these indicators, and which prior (LASSO versus regularized horseshoe) is to be preferred. For both Bayesian priors, we used default values proposed in prior literature, see Table 1. Default values for the LASSO prior were based on Park and Casella (2008), and default values for the regularized horseshoe prior were based on Piironen and Vehtari (2017a). All analysis code is available in a version-controlled repository at https://github.com/cjvanlissa/pema.

Performance indicators

Our primary performance indicator was predictive performance, a measure of model generalizability. To compute it, for each iteration of the simulation, both a training data set and a testing data set are generated from the same known population model. The number of cases in the training data vary according to the design factors of the simulation study. The number of cases in the testing data set was always 100. The models under evaluation (BRMA, RMA) were estimated on the training data, and used to predict cases in the testing data. Predictive performance was operationalized as the model's explained variance in the testing data, R^2_{test}, calculated as follows:

$$R^2_{test} = 1 - \frac{\sum_{i=1}^{k}(y_{i,test} - \hat{y}_{i,test})^2}{\sum_{i=1}^{k}(y_{i,test} - \bar{y}_{train})^2}$$

Where k is the number of studies in the testing data set, $\hat{y}_{i-test}$ is the predicted effect size for study i, and $\bar{y}_{train}$ is the mean of the training data. The R^2_{test} differs from the familiar R^2 metric: R^2 describes the proportion of variance a model explains in the training data, and it always increases as the model becomes more complex. By contrast, R^2_{test} reflects the explained variance in the testing data. Recall that BRMA was developed to reduce the risk of overfitting meta-regression models. The R^2_{test} is a useful metric to detect overfitting, which causes it to decrease, or even become negative.

The algorithm's ability to perform variable selection was evaluated by estimating sensitivity and specificity. Sensitivity P is the ability to select true positives, or the probability that a variable is selected, $S = 1$, given that it has a non-zero population effect: $P = p(S = 1||\beta| > 0)$. Specificity is the ability to identify true negatives, or the probability that a variable is not selected given that it has a zero population effect: $N = p(S = 0|\beta = 0)$.

The ability to recover population parameters β and τ^2 was examined in terms of bias and variance of these estimates. The bias is given by the deviation of the estimate from the known population value, and the variance is given by the variance of this deviation across replications of the same simulation conditions.

Design factors

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 $\tau^2 \in (.01, .04, .10)$. 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. All unique combinations of these design factors produced 1944 unique conditions. For each simulation condition, 100 data sets were generated. In each data set, the observed effect size y_i was simulated as a standardized mean difference (SMD), sampled from a non-central t-distribution.

Results

Missing data

Any iterative algorithm is susceptible to convergence problems. In such cases, the BRMA algorithms provide warning messages, but still return samples from the posterior. We were thus able to use all iterations of the BRMA algorithms, although some may have failed to converge, which would negatively impact BRMA's performance. When the RMA algorithm fails to converge, however, it terminates with an error. The RMA algorithm failed to converge in 10 replications, all characterized by low number of cases ($k \leq 40$) and high effect sizes $\beta \geq .5$. They were omitted from further analysis.

Predictive performance

Within data sets, the BRMA with a horseshoe prior had the highest predictive performance R^2_{test} 50% of the time, followed by RMA, 37%, and finally BRMA with a LASSO prior, 13%. Across data sets, the average R^2_{test} was highest for BRMA with a horseshoe prior and lowest for RMA, see Table 2. This difference was driven in part by the fact that explained variance was somewhat higher for the BRMA models when the true effect was non-zero (i.e., in the presence of a population effect), and by the fact that RMA had larger negative explained variance when the true effect was equal to zero (i.e., there was no population effect to detect).

The effect of the design factors on R^2_{test} was evaluated using ANOVAs. Note that p-values are likely not informative due to the large sample size and violation of the assumptions of normality and homoscedasticity. The results should therefore be interpreted as descriptive, not inferential, statistics. Table 3 reports the effect size η^2 of simulation conditions on R^2_{test}.

To test our research questions, we computed interactions of algorithm (HS vs. LASSO, HS vs. RMA and LASSO vs. RMA) with the other design factors. The η^2 of these differences between algorithms are also displayed in Table 3. Note that η^2 for the comparison between HS and LASSO was zero in the second decimal for all conditions; thus, this comparison was omitted from the Table. The effect of design factors by algorithm is displayed in Figure 2; these plots have been ranked from largest difference between BRMA and RMA to smallest. Results indicate that the largest differences between algorithms were due to the effect size β, number of irrelevant moderators M, and the number of cases in the training data k. Evidently, predictive performance increased most for the HS algorithm when the effect size increased above zero. As previously noted, predictive performance of RMA was the most negative when the effect size was zero. This means that RMA's explained variance in new data was below zero, a clear indication of overfitting. The HS algorithm furthermore had the consistently highest predictive performance regardless of number of irrelevant moderators or number of cases in the training data, and was relatively less affected by increases in the number of irrelevant moderators (panel b) or in the number of training cases (panel c). Conversely, RMA had relatively poor predictive performance on average, and was more responsive to increases in the number of training cases and irrelevant moderators.

Variable selection

To determine the extent to which the algorithms could perform variable selection correctly, we calculated sensitivity P, the ability to detect a true population effect, and specificity N, the ability to correctly estimate a null-effect at zero. We used all simulation conditions with $\beta > 0$, such that the population effect of the first moderator was always positive and that of the second moderator was always zero, and calculated P from the effect of the first moderator, and N from the effect of the second moderator. Finally, we computed overall accuracy as $Acc = (P + N)/2$, which reflects the trade off between sensitivity and specificity.

As the regularizing algorithms shrink all coefficients towards zero, it is unsurprising that sensitivity was highest for RMA, followed by HS and LASSO, $P_{RMA} = 0.95$, $P_{HS} = 0.91$, $P_{LASSO} = 0.89$. By contrast, specificity was higher for the regularizing algorithms, $N_{HS} = 0.98$, $N_{LASSO} = 0.97$, $N_{RMA} = 0.94$. Overall accuracy was approximately equal for RMA and HS, and was lower for LASSO, $Acc_{RMA} = 0.95$, $Acc_{HS} = 0.95$, $Acc_{LASSO} = 0.93$.

Cramer's V, an effect size for categorical variables, was used to examine the effect of design factors on sensitivity (Table 4, Figure 3) and specificity (Table 5, Figure 4). We also computed this effect size for the difference between algorithms in the number of true positives by design factor. Differences in sensitivity between the algorithms were near-zero for HS and LASSO. The difference between the two BRMA algorithms and RMA were largest for the design factor effect size β, followed by the model and number of studies k. For specificity, differences in sensitivity between HS and LASSO were largest for the number of noise moderators M, followed by the effect size β, number of studies k, and residual heterogeneity τ^2. The difference between the two BRMA algorithms and RMA were largest for the design factor number of studies k, followed by the model, the number of noise moderators M, and the effect size β. Also note that the association between design factors and specificity was not monotonously positive or negative across algorithms. Instead, some design factors had opposite effects for the two BRMA algorithms versus RMA. For instance, a larger number of studies k had a negative effect on specificity for the BRMA algorithms, but a positive effect for RMA - within the context that RMA had lower specificity on average. Conversely, a greater number of noise moderators M had a positive effect on specificity for BRMA, but a negative effect for RMA.

Ability to recover population parameters

The ability to recover population parameters β and τ^2 was examined in terms of bias and variance of these estimates. If the value of the regression coefficient as estimated by one of the algorithms is $\hat{b}$, then the bias B and variance V of this estimate can be computed as $\hat{b} - \beta$, and as the variance of $\hat{b}$ across replications of the simulation for each unique combination of design factors, respectively. For the estimated regression coefficients, HS had the greatest (negative) bias across simulation conditions, $B_{HS} = -0.07$, followed by LASSO, $B_{LASSO} = -0.06$. Surprisingly, RMA also had negatively biased estimates, $B_{RMA} = -0.01$. The effect of the design factors on the bias in estimated β was evaluated using ANOVAs. Table 6 reports the effect size η^2 of simulation conditions on the bias. The skewness of moderator variables had the largest effect on bias in estimated β for all algorithms. This was mainly because the algorithms overestimated τ^2 most when the data-generating model contained cubic terms. Simulating data with a cubic model violates the model's assumption of linearity, which biases the estimated parameters. No differences between algorithms in the effect of design factors were observed.

The variance of parameter estimates cannot be calculated on a case-by-case basis. Instead, it is calculated across replications for each simulation condition. Across simulation conditions, parameters estimated via HS had the lowest variance, $V_{HS} = 0.32$, followed by LASSO, $V_{LASSO} = 0.34$, and then RMA, $V_{RMA} = 0.38$. Online Supplemental Table S1 provides an overview of the effect size of design factors on variance of the regression coefficients. Notably, the differences between algorithms are very small; the largest effect sizes were observed for the difference between HS and RMA in the effects of effect size, sample size, and model, all with $\eta^2 < 0.01$.

Across all simulation conditions, HS had the lowest bias for the residual heterogeneity τ^2, $B_{HS} = 0.38$, followed by RMA, $B_{RMA} = 0.39$, and then LASSO, $B_{LASSO} = 0.39$. Note that all algorithms yielded positively biased estimates. The effect of the design factors on the bias in τ^2 was evaluated using ANOVAs. Table 7 reports the effect size η^2 of simulation conditions on $\hat{t^2} - \tau^2$. The design factors β and model had the largest effect on bias in estimated τ^2 for all algorithms. No differences between algorithms in the effect of design factors were observed.

The variance of the residual heterogeneity was calculated across replications for each simulation condition. The LASSO estimates of τ^2 had the lowest variance, $V_{LASSO} = 1.47$, followed by HS, $V_{HS} = 1.50$, and then RMA, $V_{RMA} = 1.71$. Online Supplemental Table S2 provides an overview of the effect size of design factors on variance of the residual heterogeneity. All differences between algorithms were small, $\eta^2 \leq 0.002$.

Applied example

This example uses the `bonapersona` data, which were included in the `pema` package with permission of the author (Bonapersona et al., 2019). This meta-analysis of over 400 experiments investigated the effects of early life adversity on cognitive performance in rodents. Note that the sample is much larger than the maximum used to validate BRMA in our simulation study. As larger samples provide greater statistical power, it should also be valid for this sample. For illustrative purposes, we use a smaller subset of the more than 30 moderators. See the `pema` package documentation (help and vignettes) for further examples.

```r
# Load relevant packages
library(pema)
library(mice)
```

```
library(rstan)
library(ggplot2)
# Select data to analyze
df <- bonapersona[ , c("yi", "vi", "mTimeLength",
                       "year", "model", "ageWeek")]
# Multiple imputation for missing values
df <- mice(df)
```

First, we estimate a model with all default settings. Based on the results of the present simulation study, the regularized horseshoe is the default prior. To see all default values, open the function documentation using `?brma`. The use of a random seed makes this example reproducible:

```
fit <- brma(yi ~ ., data = df, vi = "vi", seed = 1)
```

Running `summary(fit)` returns the posterior mean, standard deviation, and quantiles of the model parameters (see Table 8). Use the posterior mean or median (50% quantile) and 95% credible interval (2.5% - 97.5%) to perform inference on model parameters. Parameters whose 95% credible interval excludes zero are marked with an asterisk. Note that Bayesian analyses do not use the frequentist notion of significance. Instead, we say that there is a 95% probability that the true population parameter lies within the interval, given the prior and observed data. In this example, there are no moderators for which the 95% CI excludes zero. The residual heterogeneity, however, does exceed zero. The `brma()` function builds upon the `rstan` package, and its output is backwards compatible. A `brma` model can be converted to a `stanfit` object via `as.stan(fit)`. This makes it possible to benefit from the many existing convenience functions for `rstan` models. For example, it is possible to get a HPDI interval for the residual heterogeneity by running `bayestestR::hdi(as.stan(fit), parameters = "tau2")`. There are also many plotting functions for `stanfit` objects; for example, one can plot the model parameters using `plot(as.stan(fit), plotfun = "dens", pars = c("Intercept", "year"))`.

Before interpreting the results, however, it is important to assess model convergence. If any indication of non-convergence is detected during estimation, a warning will be printed. The example returns the warning that there were 331 divergent transitions, and suggests increasing the number of iterations (increasing the argument `iter` beyond its default value of 2000). Divergent transitions can result in biased estimates. If the number of divergences is small and there are no further indications of non-convergence, however, the posterior distribution is often good enough to safely interpret the results. We can examine two parameter-specific indicators of convergence by ascertaining that the number of "unique" samples from the posterior `n_eff` for each parameter is sufficiently high, and that the different chains of the estimator have mixed properly, as indicated by `Rhat` close to 1. The number of effective (independent) MCMC samples should be high relative to the total number of samples (in this case, 4000, as we used 2000 iterations on a dual-core processor). If the effective sample size is less than 10% of the total, there may be a problem - which is not the case here. The `Rhat` is a version of the potential scale reduction factor, which represents the ratio of between- and within-chain variance (Gelman & Rubin, 1992). If the chains mixed well, the `Rhat` should be close to 1. Both `n_eff` and `Rhat` indicate convergence in this example. Additional convergence diagnostics are obtained by running `check_hmc_diagnostics(as.stan(fit))`. Convergence can also be assessed visually using the function `traceplot(as.stan(fit), pars = c("Intercept", "year"))`, which provides trace plots for the MCMC draws. If the model converged, the traces of the different chains should mix well (i.e., overlap) and look like "fat caterpillars".

As explained in the section on Bayesian estimation, model convergence can be aided by increasing the amount of regularization of the prior, for example, by increasing some of the `df` parameters (see code below). In this example, increasing both `df` and `df_slab` to 5 results in only 96 divergences, compared to the original 331. This can be verified by running summary(fit2). In general, it is prudent to perform similar sensitivity analyses to determine how robust the results are to different priors. For a visual inspection of the difference in posterior distributions, use the function plot_sensitivity(), see Figure 5.

```
fit2 <- brma(yi ~ ., data = df, vi = "vi",
             prior = c("df_global" = 5, "df" = 5),
             seed = 1)
plot_sensitivity(fit, fit2) + coord_cartesian(xlim = c(-3, 3))
```

Discussion

This study presented a novel algorithm to select relevant moderators that can explain heterogeneity in meta-analyses, using Bayesian shrinkage priors. A simulation study validated the performance of two versions of the BRMA algorithm, with a regularized horseshoe prior and LASSO prior, relative to state-of-the-art meta-regression with restricted maximum likelihood estimation (RMA). Our analyses examined the algorithms' predictive performance, which is a measure of generalizability, their ability to perform variable selection, and ability to recover population parameters. Our research questions were whether BRMA offers a performance advantage over RMA in terms of any of these indicators, and which prior (horseshoe versus LASSO) is to be preferred.

Results indicated that the BRMA algorithms had higher predictive performance than RMA in the presence of relevant moderators. In the absence of relevant moderators, BRMA showed less evidence of overfitting than RMA models. In these cases, RMA models had, on average, negative predictive performance, which suggests that these models generalize poorly to new data. In the presence of an increasing number of irrelevant moderators, the BRMA algorithms' predictive performance also suffered less than that of RMA. The BRMA algorithms were also more efficient, in the sense that they achieved greater predictive performance when the number of studies in the training data was low. Across all conditions, BRMA with a horseshoe prior achieved the highest average predictive performance, and within each data set, BRMA with a horseshoe prior most often had the best predictive performance (in 50% of replications). Based on these findings, we would recommend using BRMA with a horseshoe prior when the goal is to obtain findings that generalize to new data.

With regard to variable selection, on the one hand, results indicated that the penalized BRMA algorithms had lower sensitivity: they were less able to select relevant moderators than RMA. On the other hand, the BRMA algorithms had higher specificity: they were better able to reject irrelevant moderators than RMA. Importantly, the overall accuracy was approximately equal for RMA and BRMA with a horseshoe prior. This means that the total number of Type I and Type II errors will be approximately the same when choosing between these two methods - but there is a tradeoff between sensitivity and specificity. Applied researchers must consider which is more important in the context of their research. When meta-analyzing a heterogeneous body of literature with many between-study differences, BRMA may be preferred due to its greater ability to exclude irrelevant moderators. Conversely, when meta-analyzing a highly curated body of literature with a small number of theoretically relevant moderators, RMA might be preferred.

With regard to the algorithms' ability to recover population effect sizes of moderators, we observed that BRMA with a horseshoe prior had the greatest bias towards zero across simulation conditions, followed by LASSO, and then RMA. Surprisingly, all algorithms - including RMA - provided, on average, negatively biased estimates. The variance of the estimates followed the opposite pattern, which illustrates the bias-variance trade-off. With regard to residual heterogeneity, BRMA with a horseshoe prior had the lowest bias. The BRMA algorithms also had lower variance. This suggests that the penalized regression coefficients do not compromise the estimation of residual heterogeneity. Future research might investigate under what conditions residual heterogeneity is estimated more accurately in a penalized model than in an unpenalized model. Together, these results suggest that BRMA has superior predictive performance and specificity, and provides relatively unbiased estimates of residual heterogeneity, relative to RMA.

We examined the effect of violations of the assumption of linearity by simulating data from a cubic model. In applied research, the true shape of the association between a moderator and effect size is typically unknown. Thus, model misspecification is likely to occur. One advantage of BRMA is that it can accommodate more moderators than RMA and has superior specificity. This allows researchers to specify a more flexible model to account for potential misspecification, with less concern for overfitting and nonconvergence. For example, researchers could add polynomials of continuous variables with suspected non-linear effects, or interactions between predictors. Another possible solution is to resort to non-parametric methods like random forest meta-analysis, which intrinsically accommodates non-linear effects and interactions (Van Lissa, 2020).

All simulations were conducted with default settings for the model's prior distributions, based on prior research (Park & Casella, 2008; Piironen & Vehtari, 2017b). 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 a 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.

Strengths and future directions

The present paper has several strengths. First, we included a wide range of simulation conditions, including conditions that violated the assumptions of linearity and normality. Across all conditions, BRMA displayed superior predictive performance and specificity compared to RMA. Another strength is that the present simulation study used realistic estimates of τ^2, based on data from 705 published psychological meta-analyses (Van Erp, Verhagen, Grasman, & Wagenmakers, 2017). Another strength is that the BRMA algorithms have been implemented in FAIR software (Findable, Accessible, Interoperable and Reusable): the R-package is published on the “Comprehensive R Archive Network”, and the source code is hosted on GitHub. Thanks to the use of compiled code, the BRMA algorithm is computationally relatively inexpensive.

Several limitations remain to be addressed in future research, however. One limitation is that, by necessity, computational resources and journal space limit the number of conditions that could be considered in the simulation study. To facilitate further exploration and follow-up research, all simulation data and analysis code are available online. This code can also be adapted to conduct Monte Carlo power analyses for applied research. A second limitation is that the present study did not examine the effect of multicollinear predictors. Regularizing estimators ought to have an advantage over OLS regression in the presence of multicollinearity (Hilt & Seegrist, 1977); future research ought to examine whether this advantage extends to BRMA. A third limitation is that the present study did not examine the effect of dependent data (e.g., multiple effect sizes per study). The BRMA algorithm can accommodate dependent data by means of three-level multilevel analysis. To our knowledge, there is no reason to expect that dependent data would result in a different pattern of findings than we found for independent data, but future research is required to ascertain this. A final limitation of the current implementation is that it relies on 95% credible intervals to select relevant moderators. However, these marginal credible intervals can behave differently compared to the joint credible intervals (Piironen, Betancourt, Simpson, & Vehtari, 2017). A future direction of research is therefore to implement more advanced selection procedures, such as projective predictive variable selection (Piironen & Vehtari, 2017a). Another direction for future research is the specification of different priors, aside from the horseshoe and LASSO priors that were examined in this study. A final disadvantage is that Bayesian estimation is typically more computationally expensive than frequentist estimation. One future direction of research is thus to develop a frequentist estimator for regularized meta-regression.

Recommendations for applied research

Before conducting meta-regression, researchers should be aware of its limitations (see S. G. Thompson & Higgins, 2002). These can be subdivided into four categories: 1) the curse of dimensionality and its corrolary implications for multicollinearity; 2) the ecological fallacy; 3) limited information on moderator variables, including missing data and restrictions of range. BRMA seeks to address the first of these limitations, because the problems that arise from meta-analyzing small and heterogeneous bodies of literature are so ubiquitous that they have been referred to as the primary pitfall in meta-regression (S. G. Thompson & Higgins, 2002). Nonetheless, all applicable limitations should be acknowledged in the resulting publication.

With regard to the planning and design of a BRMA meta-analysis, consider explicitly mentioning the intended use of BRMA in a preregistered analysis protocol - either as primary analysis technique, or as a contingency in the case of model non-convergence or multicollinearity. Note that BRMA is suitable for both confirmatory hypothesis tests and for exploratory analyses to ensure that no important effects were missed. Both approaches can be included in a preregistration (see Van Lissa, 2022). With regard to data extraction, it is important to strike a balance between inclusiveness and selectivity when coding moderators (S. G. Thompson & Higgins, 2002). Moderators may include theoretically relevant factors and methodological ones, such as sample demographics, methods, instruments, study quality, and publication type. A key challenge is that moderators may not always be reported. The best way to handle missing data is by recovering the relevant information by contacting authors or comparing different publications on the same data. If data remains missing, users can use multiple imputation, which is a best practice for handling missingness (see Applied example). Finally, effect sizes and their variances must be computed using suitable methods; many of which are available in the R package `metafor` (Viechtbauer, 2010).

In the Introduction, researchers should substantiate the decision to explore heterogeneity. One valid reason is prima facie heterogeneity of the body of research (Higgins et al., 2009). Another reason is the presence of theoretically relevant moderators (S. G. Thompson & Higgins, 2002). Less convincing is the practice of exploring heterogeneity only when τ^2 is significant, for two reasons: Firstly, because data-driven analysis decisions increase the risk of spurious findings (Zhang, 1992). Secondly, because tests for heterogeneity are often underpowered when the number of studies is low, and overpowered when it is high, thus limiting their usefulness (Higgins & Thompson, 2002).

With regard to data analysis, our simulation study indicates that a horseshoe prior is a suitable default. Before interpreting model parameters, one must ascertain that the algorithm has converged. Additionally, authors may consider performing a sensitivity analysis to examine whether findings are robust to different prior specifications. With regard to reporting results, researchers should report both the estimated effect of moderators and residual heterogeneity. In interpreting the regression coefficients, it should be explicitly acknowledged that regularization was used, and the parameters may thus be biased. The use of standardization and inclusion of an intercept should be reported and substantiated. As BRMA is a Bayesian method, inference is based on probability intervals instead of p-values. The null hypothesis is rejected if such intervals exclude zero. The present study compared credible intervals and HDPI intervals. Both performed identically for inference on regression coefficients. By default, `brma()` reports credible intervals - but HDPI intervals might be preferable for residual heterogeneity (which has a non-normal posterior distribution).

With regard to publication, we highly recommend making the data and code for the meta-analysis publicly available. One way to do this is by creating a reproducible research repository, for example, using the Workflow for Reproducible Code in Science (WORCS, Van Lissa et al., 2021). Transparency allows readers and reviewers to verify that methods were correctly applied, which bolsters confidence in the results. Others can easily perform sensitivity analyses by changing the analysis code. Sharing data allows the meta-analysis to be updated in the future, which increases its reuse value. Finally, sharing the model object (or code to reproduce it) allows others to obtain predictions for the expected effect size of a new study on the same topic. These predictions can be used to conduct power analysis for future research. To this end, researchers can simply enter their planned design (or several alternative designs) as new lines of data, using the codebook of the original meta-analysis, and use the published BRMA model to calculate the predicted effect size for a study with these specifications.

BRMA may not be the best solution for every situation. Several trade-offs must be considered to decide what method is most appropriate. Firstly, BRMA has higher predictive performance than RMA, which implies that it is more suitable when a researcher intends to generalize beyond the sample at hand. Conversely, RMA is more suitable when the goal is to describe the sample at hand in an unbiased manner, with less concern for generalizability to future studies. Secondly, BRMA trades off higher specificity for lower sensitivity compared to RMA, which suggests that it is more suitable when a researcher seeks to eliminate irrelevant moderators, at the cost of an increased Type II error rate. RMA might be more suitable when the researcher seeks to identify relevant moderators, at the cost of a greater Type I error rate. If many moderators are expected to be irrelevant, then BRMA may thus be preferable. Thirdly, there may be pragmatic reasons for preferring BRMA over RMA. For example, if a data set is small, or the number of moderators is high relative to the number of cases, RMA models may be empirically under-identified. This can result in convergence problems. In such cases, Bayesian estimation may converge on a solution where frequentist estimation does not (Kohli, Hughes, Wang, Zopluoglu, & Davison, 2015). Similarly, BRMA may perform better in the presence of multicollinearity among predictors, which can be examined using the function `vif()` in the R-package `metafor`. Values exceeding 5 are cause for concern. Multicollinearity increases the variance of regression coefficients. BRMA may have an advantage here, because the regularizing priors restrict variance. If multicollinearity is observed or suspected, BRMA might be preferred.

Conclusion

The present research has demonstrated that BRMA is a powerful tool for exploring heterogeneity in meta-analysis, with a number of advantages over classic RMA. BRMA had better predictive performance than RMA, which indicates that results from BRMA analysis generalize better to new data. This predictive performance advantage was especially pronounced when training data were as small as 20 studies. This is appealing because many meta-analyses have small sample sizes. BRMA further has greater specificity in rejecting irrelevant moderators from a larger set of potential candidates, while maintaining an overall variable selection accuracy equivalent to RMA. Although the estimated regression coefficients are biased towards zero by design, the estimated residual heterogeneity did not show evidence of bias in our simulation. A final advantage of BRMA over other variable selection methods for meta-analysis is that it is an extension of the linear model. Most applied researchers are familiar with the linear model, and it can easily accommodate predictor variables of any measurement level, interaction terms, and non-linear effects. Adoption of this new method is facilitated by the availability of user-friendly software in `R` and `JASP`.

Highlights

- Many applied meta-analyses concern heterogeneous bodies of literature, with many between-studies differences (moderators).
- Simultaneously, meta-analytic samples are often small. There is thus limited statistical power to account for moderators.
- The present study introduces Bayesian Regularized Meta-Analysis (BRMA), an algorithm that applies regularization to identify relevant moderators from a larger number of candidates.
- The algorithm is available via the R-package `pema` on CRAN, and via a user-friendly graphical interface in JASP.
- Readers across fields can use this method to account for between-studies heterogeneity in meta-analysis, without concern that models may be underfit or underpowered.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education*, *4*(2), 202.

Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., & From the Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice*, *63*(10), 1426–1434. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Bonapersona, V., Kentrop, J., Van Lissa, C. J., van der Veen, R., Joëls, M., & Sarabdjitsingh, R. A. (2019). The behavioral phenotype of early life adversity: A 3-level meta-analysis of rodent studies. *Neuroscience & Biobehavioral Reviews*, *102*, 299–307. https://doi.org/10.1016/j.neubiorev.2019.04.021

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

Chiquet, J., Grandvalet, Y., & Rigaill, G. (2016). On coding effects in regularized categorical regression. *Statistical Modelling*, *16*(3), 228–237.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and Behavioral Statistics*, *40*(2), 136–157. https://doi.org/10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071–4089. https://doi.org/10.1002/sim.5821

Detmer, F. J., Cebral, J., & Slawski, M. (2020). A note on coding and standardization of categorical variables in (sparse) group lasso regression. *Journal of Statistical Planning and Inference*, *206*, 1–11.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second). New York: Springer.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and Random-effects Models in Meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Hilt, D. E., & Seegrist, D. W. (1977). *Ridge, a computer program for calculating ridge regression estimates*. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.

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.

JASP Team. (2022). *JASP (version 0.16.4)[Computer software].*

Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. (2015). Fitting a Linear-Linear Piecewise Growth Mixture Model With Unknown Knots: A Comparison of Two Common Approaches to Inference. *Psychological Methods*, *20*, 259–275.

Kruschke, J. K. (2015). Chapter 12 - Bayesian Approaches to Testing a Point ("Null") Hypothesis. In J. K. Kruschke (Ed.), *Doing Bayesian Data Analysis (Second Edition)* (pp. 335–358). Boston: Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.00012-X

Lee, S. (2015). A note on standardization in penalized regressions. *Journal of the Korean Data and Information Science Society*, *26*(2), 505–516.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 30–48. https://doi.org/10.1111/bmsp.12002

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (Second). Boca Raton, FL: CRC Press.

Panityakul, T., Bumrungsup, C., & Knapp, G. (2013). On Estimating Residual Heterogeneity in Random-Effects Meta-Regression: A Comparative Study. *Journal of Statistical Theory and Applications*, *12*(3), 253–265. https://doi.org/10.2991/jsta.2013.12.3.4

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

https://doi.org/10.1198/016214508000000337

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Piironen, J., Betancourt, M., Simpson, D., & Vehtari, A. (2017). Contributed comment on article by van der Pas, Szabó, and van der Vaart. *Bayesian Analysis*, *12*(4), 1264–1266.

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b). 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

Requia, W. J., Adams, M. D., Arain, A., Papatheodorou, S., Koutrakis, P., & Mahmoud, M. (2018). Global Association of Air Pollution and Cardiorespiratory Diseases: A Systematic Review, Meta-Analysis, and Investigation of Modifier Variables. *American Journal of Public Health*, *108*(S2), S123–S130. https://doi.org/10.2105/AJPH.2017.303839

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549. https://doi.org/10.1136/bmj.d549

Rosettie, K. L., Joffe, J. N., Sparks, G. W., Aravkin, A., Chen, S., Compton, K., … Murray, C. J. L. (2021). Cost-effectiveness of HPV vaccination in 195 countries: A meta-regression analysis. *PLOS ONE*, *16*(12), e0260808. https://doi.org/10.1371/journal.pone.0260808

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., … Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, *12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

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

Sebri, M., & Dachraoui, H. (2021). Natural resources and income inequality: A meta-analytic review. *Resources Policy*, *74*, 102315. https://doi.org/10.1016/j.resourpol.2021.102315

Stan Development Team. (2022). *RStan: The R interface to Stan.*

Team, R. C. (2022). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Thompson, C. G., & Becker, B. J. (2020). A group-specific prior distribution for effect-size heterogeneity in meta-analysis. *Behavior Research Methods*, *52*(5), 2020–2030. https://doi.org/10.3758/s13428-020-01382-8

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573. https://doi.org/10.1002/sim.1187

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, *16*(4), 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

van Erp, S., 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

Van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990–2013. *Journal of Open Psychology Data*, *5*(1).

Van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using MetaForest. In R. Van De Schoot & M. Miočević (Eds.), *Small Sample Size Solutions (Open Access): A Guide for Applied Researchers and Practitioners.* CRC Press.

Van Lissa, C. J. (2022). Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Religion, Brain & Behavior*, *0*(0), 1–5. https://doi.org/10.1080/2153599X.2022.2070254

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A.-L., Peikert, A., Struiksma, M. E., & Vreede, B. M. I. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, *4*(1), 29–49. https://doi.org/10.3233/DS-210031

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wissmann, M., Toutenburg, H., et al. (2007). *Role of categorical variables in multicollinearity in the linear regression model.*

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, *79*(4), 741–746. https://doi.org/10.1093/biomet/79.4.741

Table 1

Prior parameters and corresponding arguments, along with their default values and the effect of increasing their values.

Prior	Parameter	Argument	Default	Effect
lasso	λ	`df`	1	More probability of extreme values
lasso	s	`scale`	1	Increases scale of prior
hs	λ_0^2	`df_global`	1	Increases scale of prior
hs	ν_1	`global_df`	1	Less probability of extreme values
hs	ν_2	`df`	1	Less probability of extreme values
hs	ν_3	`df_slab`	4	Less probability of extreme values
hs	s	`scale_slab`	2	Increases scale of finite slab
hs		`relevant_pars`	NULL	Increases `scale_global`

Table 2

Mean and SD of predictive R2 for BRMA with a horseshoe (HS) and LASSO prior, and for RMA, for models with a true effect (ES != 0) and without (ES = 0).

	$\bar{R^2}_{HS}$	CI_{95}	$\bar{R^2}_{LASSO}$	CI_{95}	$\bar{R^2}_{RMA}$	CI_{95}
Overall	0.42	[-0.03, 0.87]	0.42	[-0.01, 0.87]	0.39	[-0.30, 0.87]
ES = 0	0.57	[0.04, 0.89]	0.56	[0.03, 0.88]	0.55	[-0.01, 0.88]
ES != 0	-0.01	[-0.04, -0.00]	-0.01	[-0.02, 0.00]	-0.10	[-0.40, -0.01]

Table 3

Effect size of design factors on predictive R2 of the different algorithms, and of the difference between algorithms. Interpretation indicates whether a main effect was uniformly positive or negative across all algorithms.

Factor	HS	LASSO	RMA	HS vs. LASSO	HS vs. RMA	LASSO vs. RMA	Interpretation
ω	0.02	0.01	0.01	0.00	0.00	0.00	negative
β	0.77	0.76	0.70	0.00	0.01	0.02	positive
k	0.02	0.02	0.06	0.00	0.01	0.01	positive
n	0.05	0.05	0.02	0.00	0.00	0.00	positive
Model	0.17	0.17	0.11	0.00	0.00	0.00	positive
M	0.00	0.00	0.04	0.00	0.01	0.01	negative
τ^2	0.05	0.05	0.03	0.00	0.00	0.00	negative

Table 4

Effect size (Cramer's V) of design factors, and of the difference between algorithms, on sensitivity (P).

Factor	P_{HS}	P_{LASSO}	P_{RMA}	$P_{HSvs.LASSO}$	$P_{HSvs.RMA}$	$P_{LASSOvs.RMA}$	Interpretation
k	0.21	0.23	0.17	0.01	0.02	0.02	positive
n	0.08	0.09	0.07	0.00	0.01	0.01	positive
β	0.36	0.37	0.28	0.01	0.04	0.04	positive
τ^2	0.10	0.10	0.08	0.00	0.01	0.01	negative
ω	0.09	0.10	0.08	0.00	0.01	0.01	negative
M	0.05	0.05	0.02	0.00	0.01	0.01	negative
Model	0.31	0.33	0.22	0.01	0.03	0.03	positive

Table 5

Effect size (Cramer's V) of design factors, and of the difference between algorithms, on specificity (N).

Factor	N_{HS}	N_{LASSO}	N_{RMA}	$N_{HSvs.LASSO}$	$N_{HSvs.RMA}$	$N_{LASSOvs.RMA}$	Interpretation
k	0.02	0.03	0.02	0.03	0.13	0.13	other
n	0.00	0.01	0.00	0.01	0.02	0.02	other
β	0.01	0.02	0.01	0.03	0.06	0.06	other
τ^2	0.02	0.01	0.02	0.03	0.01	0.01	other
ω	0.00	0.01	0.00	0.01	0.02	0.02	other
M	0.04	0.03	0.01	0.11	0.08	0.08	other
Model	0.02	0.03	0.01	0.01	0.08	0.08	positive

Table 6

Effect size of design factors on bias in beta squared for the different algorithms, and of the difference between algorithms.

Factor	HS	LASSO	RMA	HS vs. LASSO	HS vs. RMA	LASSO vs. RMA
ω	0.16	0.15	0.15	0.00	0.00	0.00
β	0.01	0.00	0.00	0.00	0.00	0.00
k	0.00	0.00	0.00	0.00	0.00	0.00
n	0.02	0.02	0.01	0.00	0.00	0.00
Model	0.01	0.00	0.00	0.00	0.00	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00
τ^2	0.00	0.00	0.00	0.00	0.00	0.00

Table 7

Effect size of design factors on bias in tau squared for the different algorithms, and of the difference between algorithms.

Factor	HS	LASSO	RMA	HS vs. LASSO	HS vs. RMA	LASSO vs. RMA
ω	0.01	0.01	0.00	0.00	0.00	0.00
β	0.12	0.13	0.11	0.00	0.00	0.00
k	0.00	0.00	0.00	0.00	0.00	0.00
n	0.01	0.01	0.01	0.00	0.00	0.00
Model	0.11	0.12	0.10	0.00	0.00	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00
τ^2	0.00	0.00	0.00	0.00	0.00	0.00

Table 8

Summary of model parameters for the applied example.

	mean	sd	2.5%	50%	97.5%	n_eff	Rhat
Intercept	-27.34	16.63	-59.46	-27.42	1.54	1,188.50	1.00
mTimeLength	0.00	0.01	-0.02	0.00	0.00	188.60	1.03
year	0.01	0.01	0.00	0.01	0.03	1,187.99	1.00
modelLG	0.13	0.16	-0.09	0.09	0.52	998.39	1.00
modelLNB	0.15	0.13	-0.03	0.14	0.43	544.65	1.01
modelM	0.05	0.07	-0.04	0.03	0.21	586.43	1.00
modelMD	0.04	0.09	-0.14	0.02	0.26	264.84	1.02
ageWeek	-0.01	0.01	-0.02	0.00	0.00	446.33	1.01

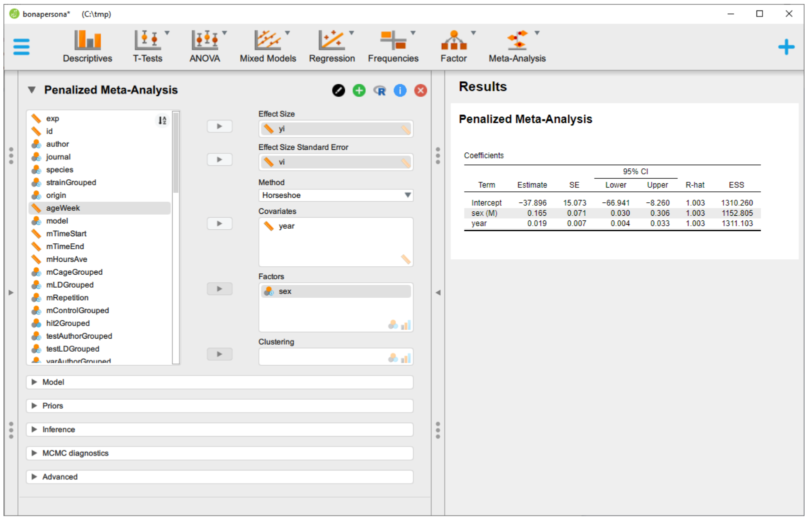

Figure 1. Using BRMA via the JASP software package.

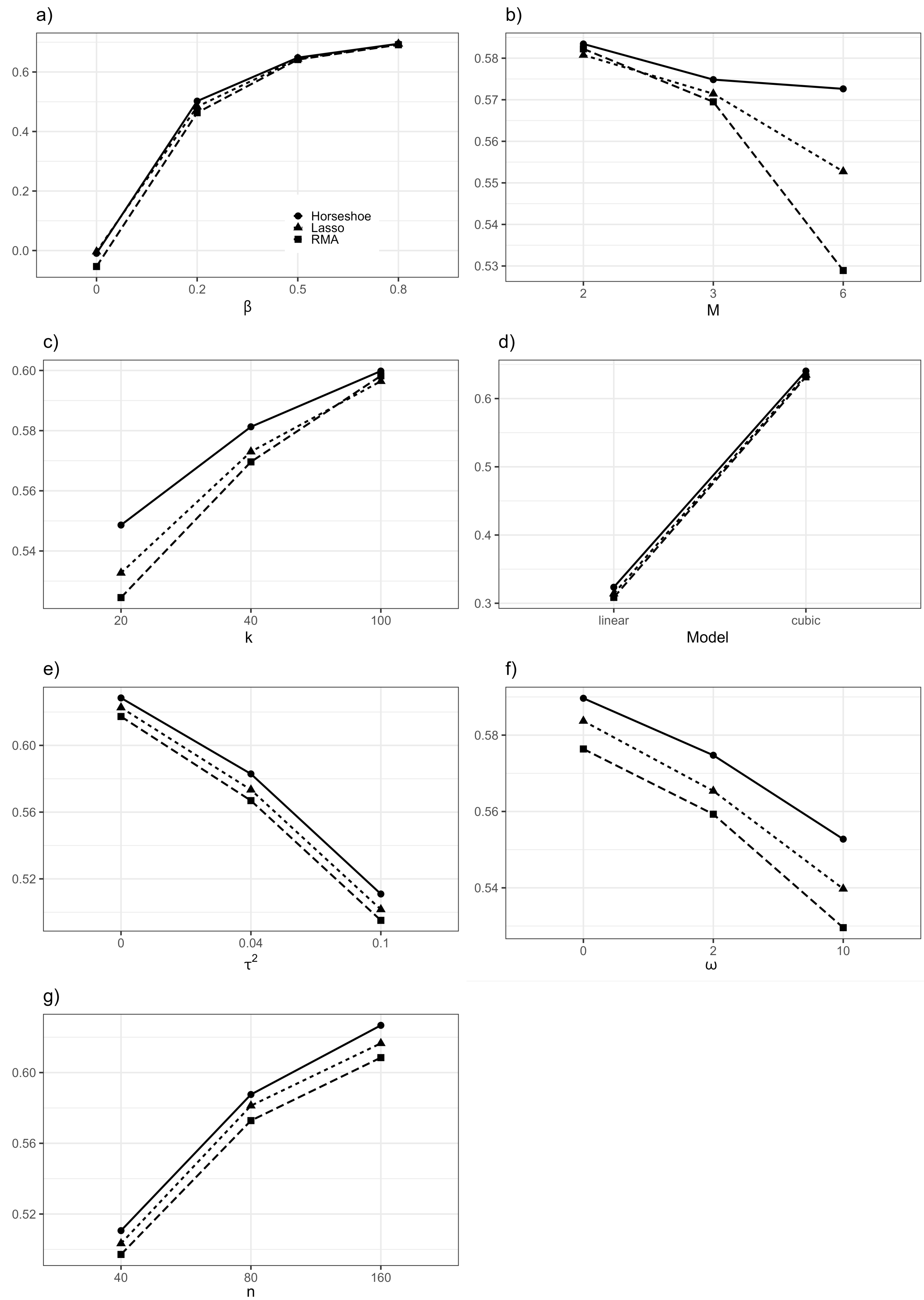

Figure 2. Predictive R2 for BRMA with horseshoe (HS) and LASSO prior, and RMA. Plots are sorted by largest performance difference between BRMA and RMA.

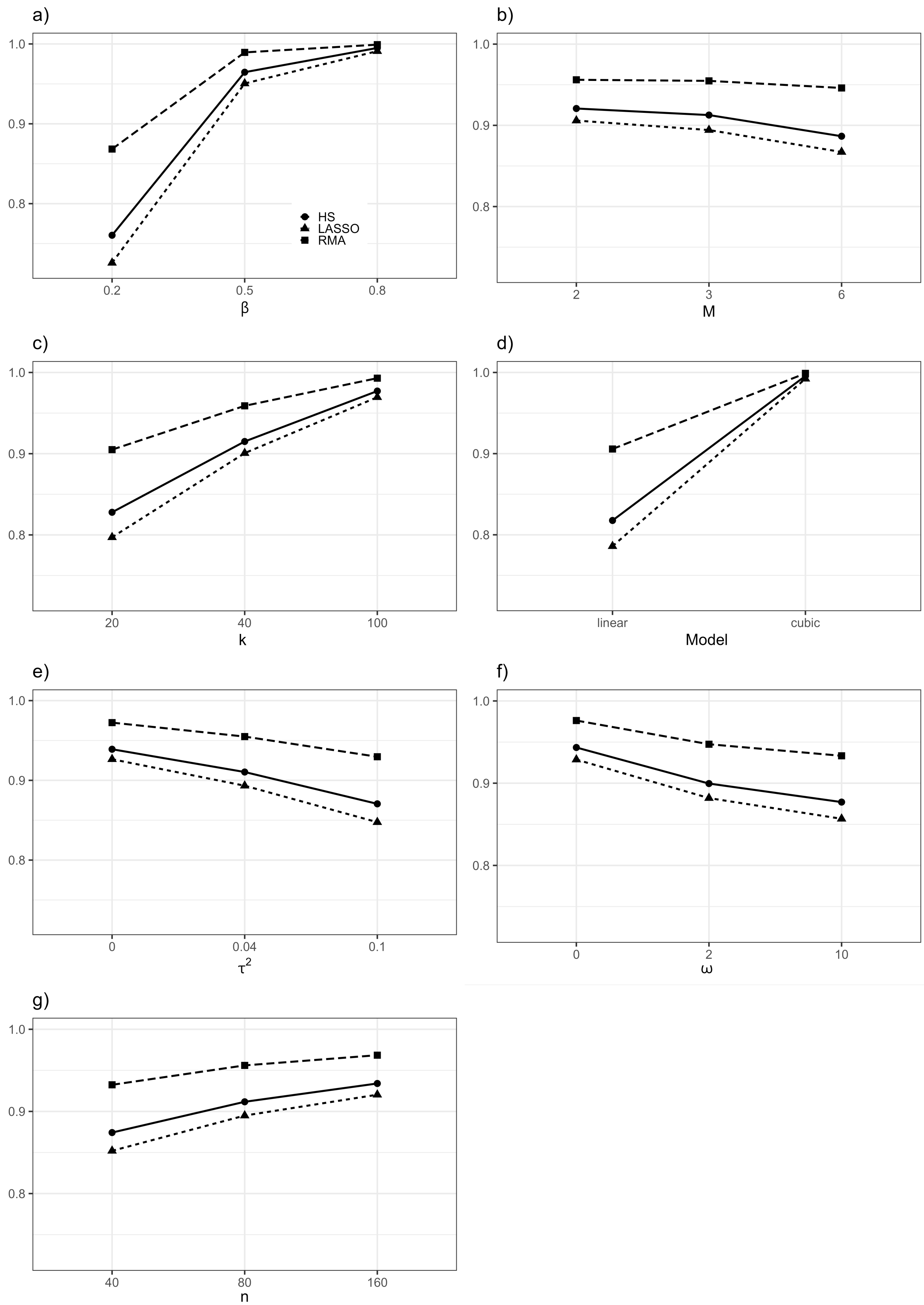

Figure 3. Sensitivity by design factors for the HS (circle, solid line), LASSO(triangle, dotted line) and RMA (square, dashed line) algorithms.

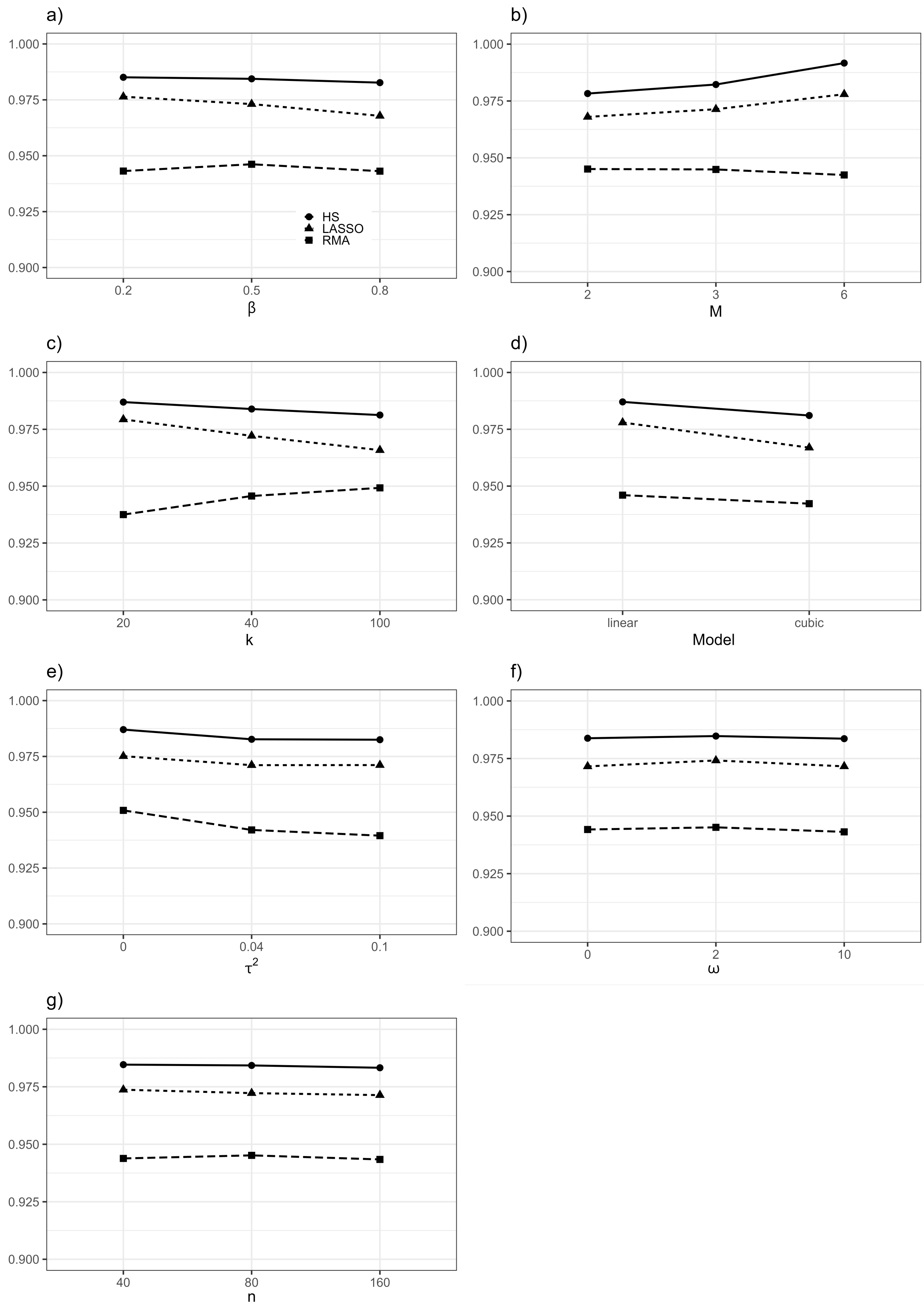

Figure 4. Specificity by design factors for the HS (circle, solid line), LASSO(triangle, dotted line) and RMA (square, dashed line) algorithms.

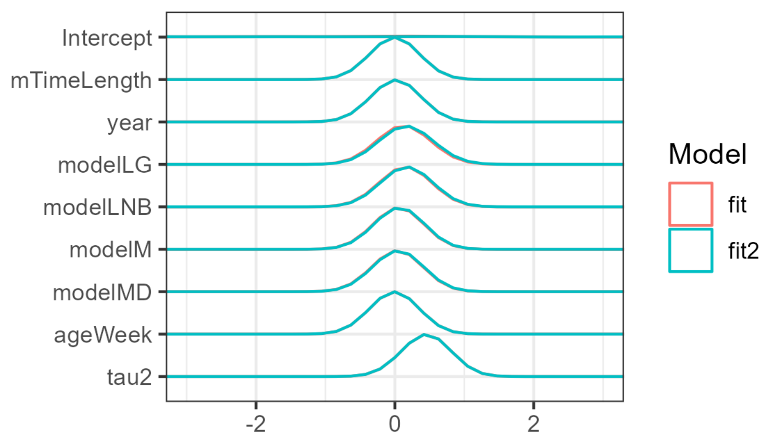

Figure 5. A rudimentary visual check for sensitivity to different priors.