Select relevant moderators in meta-regression using Bayesian penalization

Caspar J. Van Lissa[1,2] & Sara van Erp[1]

[1] Utrecht University, dept. Methodology & Statistics

[2] Open Science Community Utrecht

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Caspar J. Van Lissa: Conceptualization, Writing - Original Draft Preparation, Programming front-end; Sara van Erp: Writing - Contributions and feedback, Programming back-end.

Correspondence concerning this article should be addressed to Caspar J. Van Lissa, Padualaan 14, 3584CH Utrecht, The Netherlands. E-mail: c.j.vanlissa@uu.nl

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

Select relevant moderators in meta-regression using Bayesian penalization

Skeleton lasso/pema paper 1.) What is Meta-analysis? 2.) What is meta-regression and how does it complement meta-analysis? a.) Introduce Moderators b.) Study Heterogeneity + Random Sampling Error (and their difference)

2.5.) Fixed vs. random effects a.) Shortcomings of fixed effect models 3.) Shortcomings of current meta regressions w.r.t. estimating coefficients and heterogeneity: a.) Small sample size / overfitting b.) Non-normal data 4.) Various methods to estimate heterogeneity (and coefficients) a.) The use of WLS and REML 5.) Intro to Frequentist linear methods/Bayesian methods and Random forests, along with their (dis-)advantages: a.) Rma: uses WLS for estimation b.) MetaForest (Random Effects): Uses Random Forest Algorithm c.) Lasso Pema: uses penalized Lasso d.) Horseshoe Pema: uses horseshoe priors 6.) Goal of the current study 7.) Means of attaining goal and evaluation of performance: a.) simulation study b.) algorithmic performance c.) design factors d.) Impact of design factors on algorithim performance e.) Hypotheses of algorithmic performances

colour coding: I colour coded the text as to know from which file the text is copied GREEN = derived from 'Thesis_Metaforest' BLUE = Thesis_lasso BLACK = internship_report RED = Inserted myself

Introduction

Meta-analysis is a quantitative form of evidence synthesis, whereby effect sizes from multiple similar studies are aggregated. In its simplest form, this aggregation consists of the computation of a summary effect as a weighted average of the observed effect sizes. This average is weighted to account for the fact that some observed effect sizes are assumed to be more informative about the underlying population effect. Each effect size is assigned a weight that determines how influential it is in calculating the summary effect. This weight is based on specific assumptions; for example, the *fixed effect* model assumes that all observed effect sizes reflect one underlying true population effect size. This assumption is well-suited to the situation where effect sizes from close replication studies are meta-analyzed (**higgins_re-evaluation_2009?**, fabrigar_conceptualizing_2016, maxwell_is_2015). The *random effects* model, by contrast, assumes that population effect sizes follow a normal distribution. Each observed effect size provides information about the mean and standard deviation of this distribution of population effect sizes. This assumption is more appropriate when studies are conceptually similar and differences between them are random (**higgins_re-evaluation_2009?**, fabrigar_conceptualizing_2016, maxwell_is_2015).

Not all heterogeneity in effect sizes is random, however. Quantifiable between-study differences may introduce systematic heterogeneity. Such between-study differences are known as "moderators." For example, if studies have been replicated in Europe and the Americas, this difference can be captured by a binary moderator called "continent." Alternatively, if studies have used different dosages of the same drug, this may be captured by a continuous moderator called "dosage." Systematic heterogeneity in the observed effect sizes can be accounted for using *meta-regression* (Viechtbauer & López-López, 2015). This technique provides estimates of the effect of one or more study characteristics on the overall effect size, as well as of the overall effect size and residual heterogeneity after controlling for their influence.

One common application of meta-analysis is to summarize existing bodies of literature. In such situations, the number of moderators is often relatively high because similar research questions have been studied in different laboratories, using different methods, instruments, and samples. Each of these between-study differences could be coded as a moderator, and some of these moderators may explain systematic heterogeneity.

It is theoretically possible to account for the influence of multiple moderators using meta-regression. However, like any regression-based approach, meta-regression requires a relatively high number of cases (studies) per parameter obtain sufficient power to examine heterogeneity. In practice the number of available studies is often too low to examine heterogeneity reliably (Riley, Higgins, & Deeks, 2011). At the same time, there are many potential sources of heterogeneity, as similar research questions are studied in different laboratories, using different methods, instruments, and samples. This leads to a problem known as the "curse of dimensionality": the number of candidate moderators is large relative to the number of cases in the data. Such cases do not fit comfortably into the classic meta-analysis paradigm, which, like any regression-based approach, requires a high number of cases per parameter. Between-studies thus presents a non-trivial challenge to data aggregation using classic meta-analytic methods. At the same time, it also offers an unexploited opportunity to learn which differences between studies have an impact on the effect size found, if adequate exploratory techniques can be developed.

Addressing the curse of dimensionality necessitates *variable selection*: the selection of a smaller subset of relevant moderators from a larger number of candidate moderators. One way to perform variable selection is by relying on theory. However, in many fields of science, theories exist at the individual level of analysis (e.g., in social science, at the level of individual people). These theories do not necessarily generalize to the study level of analysis. Using theories at the individual level for moderator selection at the study level amounts to committing the ecological fallacy: generalizing inferences across levels of analysis (**jargowskyEcologicalFallacy2004?**). To illustrate what theory at the study level of analysis might look like, consider the so-called *decline effect*. It is a phenomenon whereby effect sizes in a particular tranche of the literature seem to diminish over time (**schoolerUnpublishedResultsHide2011?**). It has been theorized that the decline effect can been attributed to regression to the mean: A finding initially draws attention from the research community because an anomalously large effect size has been published, and subsequent replications find smaller effect sizes. Based on the decline effect, we might thus expect the variable "year of publication" to be a relevant moderator of study effect sizes.

Note that this prediction is valid even if year is orthogonal to the outcome of interest within each study. Until more theory about the drivers of between-study heterogeneity is developed, however, this approach will have limited utility for variable selection.

An alternative solution is to rely on statistical methods for variable selection. This is a focal issue in the discipline of machine learning (**hastieElementsStatisticalLearning2009?**). One technique that facilitates variable selection is *regularization*: shrinking model parameters towards zero, such that only larger parameters remain. Although this technique biases the parameter estimates, it also reduces their variance, which has the advantage of producing more generalizable results that make better predictions for new data (see **hastieElementsStatisticalLearning2009?**). This paper introduces *Bayesian regularized meta-regression* (BRMA), an algorithm that uses Bayesian estimation with regularizing priors to perform variable selection in meta-analysis. The algorithm is implemented in the function `brma()` in the R-package `pema`.

Statistical underpinnings

To understand how BRMA estimates the relevant parameters and performs variable selection, it is instructional to first review the statistical underpinnings of the aforementioned classic approaches to meta-analysis. First is the fixed-effect model, which assumes that each observed effect size T_i is an estimate of an underlying true effect size Θ (**hedgesFixedRandomeffectsModels1998?**). The only cause of heterogeneity in observed effect sizes is presumed to be effect size-specific sampling variance, 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 $i =$ 1,2, . . . k) are given by:

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

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

Under the fixed effect model, the estimated population effect size $\hat{\theta}$ is obtained by computing a weighted average of the observed effect sizes. If sampling error is assumed to be the only source of variance in observed effect size, then it follows that studies with smaller standard errors estimate the underlying true effect size more precisely. The fixed-effect weights are thus 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:

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

–>

Whereas the fixed-effect model assumes that only one true population effect exists, the random-effects model assumes that true effects may vary for unknown reasons, and thus follow a (normal) distribution of their own (Hedges & Vevea, 1998). This heterogeneity of the true effects is represented by their variance, τ^2. The random effect model thus assumes that the heterogeneity in observed effects can be decomposed into sampling error and between-studies heterogeneity, resulting in the following equation for the observed effect sizes:

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

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

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

In this model, Θ is the mean of the distribution of true effect sizes, and τ^2 is its variance, which can be interpreted as the variance between studies.

If the true effect sizes follow a distribution, then even less precise studies (with larger sampling errors) may provide some information about this distribution. Like fixed-effect weights, random effects weights are still influenced by sampling error, but this influence is attenuated by the estimated variance of the true effect sizes. The random-effects weights are thus given by $w_i = \frac{1}{v_i + \hat{\tau}^2}$. It is important to note that, whereas the sampling error for each individual effect size is treated as known, between-study heterogeneity τ^2 must be estimated. This estimate is represented by $\hat{\tau}^2$.

Meta-regression. The random effects model assumes that causes of heterogeneity in the true effect sizes are unknown, and that their influence is random. Oftentimes, however, there are systematic sources of heterogeneity in true effect sizes. These between-study differences can be coded as moderators, and their influence can be estimated and controlled for using meta-regression. Meta-regression with p moderators can be expressed with the following equation, where $x_{1 \ldots p}$ represent the moderators, and $\beta_{1 \ldots p}$ the regression coefficients:

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

(8)

Note that β_0 represents the intercept of the distribution of true effect sizes after controlling for the moderators and the error term ζ_i represents residual between-studies heterogeneity. This term is still included because unexplained heterogeneity often remains after accounting for the moderators (Thompson & Sharp, 1999). This is a mixed-effects model; the intercept and effects of moderators are treated as fixed and the residual heterogeneity as random (Viechtbauer & López-López, 2015).

To solve this model, the regression coefficients and residual heterogeneity must be estimated simultaneously. Numerous methods have been proposed to estimate meta-regression models, the most commonly used of which is restricted maximum likelihood (REML). REML is an iterative method, meaning it performs the same calculations repeatedly, updating the estimated regression coefficients and residual heterogeneity until these estimates stabilize. In contrast to a regularized analysis technique, this estimator produces low bias, which means that the average value of the estimated regression coefficients and residual heterogeneity is close to their true values (Panityakul et al., 2013; Hardy & Thompson, 1996).

Regularized regression

Regularized regression biases parameter estimates towards zero by including a shrinkage penalty in the estimation process. Before examining the Bayesian case, we will explain the principle using frequentist OLS regression as an example. OLS regression estimates the model parameters by minimizing the Residual Sum of Squares (RSS) of the dependent variable, which is given by:

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

The resulting parameter estimates are those that give the best predictions of the dependent variable in the present dataset. Penalized regression, by contrast, adds a penalty term to the quantity to be minimized. One commonly used penalty is the L1-norm of the regression coefficients, or LASSO penalty, which corresponds to the sum of their absolute values. This gives the penalized residual sum of squares:

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

Where λ is a tuning parameter that determines how influential the penalty term will be. If λ is zero, the shrinkage penalty has no impact at all and the penalized regression will produce the OLS estimates. If $\lambda \to \infty$, all coefficient shrink towards zero, producing the null model. Because the penalty term is a function of the regression coefficients, the optimizer has an incentive to keep the regression coefficients as small as possible. Note that theLASSO penalty is but one example of a shrinkage penalty; other penalties exist, with some unique properties.

Bayesian estimation

Numerous methods have been proposed to accurately estimate the residual heterogeneity, including the Hedges (HE), DerSimonian–Laird/Method of Moments (DL), Sidik and Jonkman (SJ), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), and Empirical Bayes (EB) method. These methods are mostly divided into two groups: closed-form or non-iterative methods and iterative methods. The main difference between these groups is that the closed form group uses a predetermined number of steps to provide an estimation for the residual heterogeneity, whereas the iterative methods run multiple iteration, as the name suggests, to converge to a solution when a specific criterion is met. It is important to note that some iterative methods do not produce a solution when they fail to converge after a predetermined amount of iteration.

In our scenario we are especially interested in an estimator which performs well under the condition of a relative low number of studies. The Restricted Maximum Likelihood (REML) seems to produce the lowest bias under this condition and is therefore preferred (Panityakul et al., 2013; Hardy & Thompson, 1996). The REML is an iterative method and needs a starting estimation of τ^2 to start, usually it gets estimated by one of the non-iterative methods (Viechtbauer & López-López, 2015). Besides the starting value of τ^2, it needs in every iteration an estimation of the regression coefficients of the moderators. These are typically estimated by using the Weighted Least Squares (WLS) method. This is a variation of the Ordinary Least Squares (OLS), but in the case of meta-analysis it is necessary to assess weights to the coefficients. In systematic reviews large variation in standard errors is often observed, which will result in large heteroscedasticity in the estimation of the effects (Stanley & Doucouliagos, 2017). The addition of weights is a way to adjust for this heteroscedasticity. The weights are formulated as presented in equation (5).

The usage of a WLS method to estimate the regression coefficient may be problematic in the situation where a lot of moderators are measured without their specific effects, when the amount of studies is low and when moderators are dichotomous. The use of a least squares method will cause problems with the prediction accuracy and the model interpretability (James, Witten, Hastie, & Tibshirani, 2013). In the situation where a lot of moderators are measured and blindly included in the model, it may as well be the case that variables are included that are in fact not associated with the response. Including irrelevant variables in the model lowers the interpretability of the model (James et al., 2013). An approach is necessary that automatically excludes the variables that are irrelevant i.e. performs variable selection. As explained before, in meta-analysis it is often the case that the number of moderators closely approaches or even exceeds the number of studies included in the analysis. A least squares method will display a lot variability in the fit when the number of variables is not much smaller than the number of studies (James et al., 2013). This means that the least squares method over fits the data and loses its power to be generalizable to future observations. When the number of variables exceeds the number of studies, the least squares method fails to produce one unique estimate and the method should not be used at all.

However, a least squares method could still be somewhat valuable in some situations. It is extremely suitable to estimate a linear relationship. In the case of dichotomous moderators, the relationship is always perfectly linear. A powerful non-linear estimation tool is in the situation of dichotomous moderators unnecessary and would not perform better at all. Whenever a non-linear relation gets fitted on data with an underlying linear relation, it will cause problems when this fit gets used for the prediction of future data. Given the various arguments, this paper provides an approach to tackle this problem of the least squares methods whilst still making use of a linear method. The weighted least squares are replaced with the so-called LASSO regression for the estimation of the regression coefficients. This algorithm shrinks or penalizes the regression coefficients and performs variable selection (James et al., 2013; Hesterberg, Choi, Meier, & Fraley, 2008). –>

Algorithms and simulation + goal study The goal of the present study was to test whether a ME-MRA model with the lasso algorithm is able to outperform the ME-MRA with least squares regression. More specifically, if the lasso is able to outperform the least squares when in situation where the amount studies included in the analysis is fairly low. To test this, two different algorithms are used; one called the rma, which makes use of the WLS regression, and the lma, which makes use of a regularized lasso regression. To test the lma and rma algorithms on the performance criteria, a simulation study is performed. A simulation of the data is preferred over the use of real data. Simulated data can be shaped to such an extent that it will have the all desired characteristics to test the performance of the algorithm. Besides that, if simulated correctly, it will not have any systematic errors or noise due to underlying models and it is more cost efficient. Simulations are useful for evaluation of new methods like MetaForest and for the comparison with alternative methods like metaCART and the classic approaches.

Performance Criteria The algorithms are evaluated on three different performance criteria: The algorithms' predictive performance, their ability to estimate the residual heterogeneity and their ability to detect and select the right moderators. The predictive performance of the algorithms is defined by how well the algorithm is able to predict future data. The algorithms have to estimate a model on a "training" dataset and then use this model to see how well it fits on a second "testing" dataset. This is operationalized as the cross-validated R^2_{cv} (Van Lissa, 2017). The R^2_{cv} is calculated using the fraction of variance explained by the model on the testing dataset, relative to faction of variance explained by the mean of the testing dataset. The mean of the testing dataset is the best prediction for the testing data when there is no model present (van Lissa, 2017). The calculation of$R^2_{cv}$is expressed by the following equation:

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (9)$$

With n being the number of studies in the testing dataset, $\hat{y}_i$ being the estimation for study i, and $\bar{y}_i$ being the mean of the training dataset.

The ability of the algorithms to estimate the residual heterogeneity is by simply taking the value of τ^2 which the algorithm produces. The true value of the residual heterogeneity is subtracted of the estimated value, solely to make the values more interpretable. This means that a correct estimation of the residual heterogeneity will be expressed by a value which exactly or close to zero. The residual heterogeneity is used as a performance criterion because it is suspected that the lma model might not always be able to predict residual heterogeneity correctly.

Variable selection is defined in terms of the algorithms ability to accredit positive variable importance values to relevant moderators. Variable importance measures capture the relative contribution of various moderators.

Design factors In the simulation study, meta analytic datasets will be simulated. These datasets consist of two separate sub-datasets, a training- and a testing dataset. Both sub-datasets will have the same characteristics with the exception of the number of studies included. Certain characteristics of the sub-datasets will be manipulated to test how well the algorithms perform under certain conditions. For each combination of characteristics, or design factors, 100 datasets will be simulated. The design factors that will be manipulated are the number of studies in the training data k (22, 40 and 80), the average within-study sample size $\bar{n}$ (40, 100 and 200), the population effect size β (.2, .5 and .8) and the residual heterogeneity τ^2 (.01, .04 and .1). All the datasets will contain 20 moderators of which 10 are relevant and 10 are irrelevant. The moderators are binary and follow $\sim \mathcal{B}\rceil\nabla\setminus(0.5)$, which corresponds to an equal chance of being either one or zero. The dependent variable y_i represented by a $Hedges'g$. This is an estimator which takes the standardized mean difference between a treatment and control group and is commonly used in meta-analysis (Van Lissa, 2017). The true effect size θ_i is sampled out of a normal distribution. The mean is computed by the assessing the values of the coefficients β_j, with the values of the moderators and with the residual heterogeneity τ^2 (Van Lissa, 2017). This is in line with the calculation of θ_i represented in equation (6). The sampling error σ_i^2 is formed by varying the sizes of the samples of each study. The sample sizes $n_i \sim \mathcal{N}(\bar{n}, \frac{\bar{n}}{3}$ (Van Lissa, 2017).

Data were simulated using the random-effects model, based on four models: (A) Main effect of one moderator, $\mu_i = \beta_1 x_{1i}$ (B) Two-way interaction, $\mu_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$ (E) Non-linear, cubic relationship, $\mu_i = \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3$ (F) Exponential relationship, $\mu_i = \beta_1 e^{x_{1i}}$

Impact of design factors and hypotheses These design factors are chosen on purpose, because they are hypothesized to have an influence on the predictive performance of the algorithms. The effect of the design factors ought to be either positive or negative on the data. This means that some factor should, by increasing, make the data easier to be analyzed, or make it more difficult to analyze. The amount of studies included in the training data k has a positive influence on the variance explained by the different algorithms. This is due to the fact that there are simply more data points available to fit a model on. The lma algorithm should be superior on the low value of k over the rma algorithm. The effect size β has a positive impact on the ability of the algorithms to explain variance. It can be hypothesized that the lma performs better at lower values of β because it is better equipped to detect and select variables when even when the amount of signal is low. The residual heterogeneity τ^2 should have a negative influence on the interpretability of the data. Differences between the two algorithms could be present, but it remains unclear which would perform better. The lma might perform better when the amount of signal in the data is low or the noise is high, but it is also suspected to overestimate the amount of heterogeneity and this could worsen if the τ^2 increases. The $\bar{n}$ greatly influences the quality of the data. Higher values of within-study sample sizes reduce the sampling error. This will lead to a better prediction by the algorithms. In conclusion: higher values of k, β and $\bar{n}$ will increase the quality of the data, where higher values of τ^2 decrease the quality of the data. The lma is suspected to perform significantly better when the quality of the data is low, especially when the amount of studies in the sample is low, with the exception of the performance of the lma on the estimation of the residual heterogeneity

Results

There were 3888 condition combinations in total. The algorithms ran on 100 different datasets on each of those, leaving 388800 cases to analyze. There were 20 of those cases for which the RMA algorithm had missing values on all metrics. Closer inspection showed that both the cubic and exponential model each contributed ten times to the missing values and only when 2, 3 or 6 moderators were taken up in the model. However, since the missing cases make up only 0.005% of the data, the cases were chosen to be omitted from further analysis entirely. Another observation is that the two-way interaction model only had results when the number of moderators in the model was either 3, 4 or 7, while the other models only had results when there were 2, 3 or 6 moderators. It is therefor more challenging to compare and interpret the effect the moderators had on the performance criteria between the two-way interaction and the other models.

Predictive performance

Predictive performance was operationalized by calculating the R^2_{test} and MSE_{test} for every combination of design factors, in further test denoted as R^2 and MSE respectively. The densities for the R^2_{test} and MSE_{test} values were skewed however, which is why it was chosen to use the median R^2 as the metric for predictive performance, rather than the mean. The spread of the metrics was described using the Mean Absolute Deviation [MAD], rather than the standard deviation. It was found that the Horseshoe, Lasso and RMA algorithm performed similarly overall, $R^2_{Hs} = 0.51 \pm 0.36$, $MSE_{Hs} = 0.21 \pm 0.18$; $R^2_{Lasso} = 0.50 \pm 0.37$, $MSE_{Lasso} = 0.21 \pm 0.19$; $R^2_{RMA} = 0.50 \pm 0.37$, $MSE_{RMA} = 0.22 \pm 0.23$. The MetaForest algorithm performed worst on R^2: $R^2_{Mf} = 0.35 \pm 0.38$, $MSE_{Mf} = 0.22 \pm 0.19$

To determine the effect of the design factors on R^2 for all algorithms, four separate ANOVA's were performed; one per algorithm. The effect size η^2 per condition per algorithm, including for all two-way interactions can be found in table 1. Do note that the ANOVA's were performed with the normality assumption violated. The estimates serve mostly as a guidance, rather than an absolute result.

Not too surprisingly, It was found that the true effect size β had the largest effect on R^2 for all algorithms. As β increased, the performance of all algorithms increased as well. However, β did interact with the model that was estimated, being either a linear, two-way interaction, cubic or an exponential model. A graphical representation of the interaction is shown in image 1. When the exponential and cubic model were estimated, the increase of R^2 slowed for every higher value of β. For the estimation of the two-way interaction model the increase only slowed when β went from 0.5 to 0.8, while the steepest increase for the linear model estimation was when β increased from 0.2 to 0.5. Also noteworthy is that R^2 stagnated during estimation of the cubic model as β went up to 0.5, while for the other models R^2 did keep increasing. There was little difference in median R^2 between Horsehoe, Lasso and RMA, while MetaForest performed worst.

The second largest marginal effect was that of the estimated model. All algorithms had the highest R^2 under the cubic model, followed by a similar performance on the two-way interaction and exponential models. All algorithms performed worst for the linear model. There again was little difference in performance between the Pema algorithms and RMA, although MetaForest performed worst. Image 2A shows the relationship.

There also was a moderate interaction effect between the estimated model and the amount of skewness of the input data α, especially for the Pema algorithms. Again, Pema and RMA algorithms performed best, followed by MetaForest in all conditions. Most obvious to note is that all algorithms, except for when the linear model was estimated, generally performed better on more skewed data, although the algorithms did perform worse during estimation of the two-way interaction model when α went from 5 to 10. Overall performance was best during estimation of the cubic model, but the performance difference between estimated models decreased as α increased. Image 3 shows the relationship.

The true residual heterogeneity τ^2 had a negative linear relationship for all algorithms on R^2. That is, as τ^2 increased, R^2 decreased. Image 2B shows the relationship.

The mean sample size per study $\bar{n}$ also had a moderate effect. For all algorithms the effect of $\bar{n}$ was positively linearly related with R^2. Again, the Pema and RMA algorithms performed better than MetaForest. Image 2C shows the relationship.

An especially large effect was found for the number of studies used in the training data κ for MetaForest, while this effect was substantially smaller for RMA, Lasso and Horseshoe. The relationship is positively linear for all algorithms, but the slope is especially steep for MetaForest. Image 2D shows the relationship.

Finally, the number of moderators did not have a big effect for the Pema algorithms, while for RMA and MetaForest the effect was more noticeable. The relationship is shown in image 7. The relationship is generally negative with more moderators meaning worse performance, although an increase can be observed as the number of moderators increase from 4 to 6 for all algorithms except MetaForest. This increase in performance for MetaForest appears when the number of moderators go from 3 to 4.

Estimating residual heterogeneity

The ability of the algorithms to correctly estimate τ^2 was operationalized by subtracting the true value for τ^2 from the τ^2 estimated by the algorithms. Again, the median and Mean Absolute Deviation were used as metrics for performance. The RMA algorithm showed the best results, $\Delta\tau^2_{RMA} = 0.02 \pm 0.06$, followed by the MetaForest algorithm $\Delta\tau^2_{Mf} = 0.09 \pm 0.13$. The Pema algorithms performed worst $\Delta\tau^2_{Lasso} = 0.23 \pm 0.18$; $\Delta\tau^2_{Hs} = 0.23 \pm 0.17$. The finding that all medians are positive implies that all algorithms have a bigger tendency to overestimate τ^2 than underestimate it. One comment to make is that uncertainty of the estimates generally increased as $\Delta\tau^2$ also increased. This implies that there was more variation in performance as median performance worsened.

To determine the effect of the design factors on $\Delta\tau^2$ for all algorithms, four separate ANOVA's were performed. The effect size η^2 per condition per algorithm, including η^2 for all two-way interactions can be found in table 2. Again, the assumption of normality was violated.

The biggest predictor on the correct estimation of τ^2 was the estimated model. This was mainly because the algorithms overestimated τ^2 most when the model contained cubic terms. Image 5A shows the marginal relationship of the estimated model on $\Delta\tau^2$. It becomes more clear why this overestimation occurred when showing the interaction between β and the model estimated on $\Delta\tau^2$, shown in image 6. First note the general trend that during estimation of all models τ^2 got more overestimated as β increased, except during estimation of the linear model, where the effect of β on $\Delta\tau^2$ was close to zero, except for MetaForest. However, note the scales for the y-axes. While estimating the two-way interaction, linear and exponential model, $\Delta\tau^2$ stayed well within a confined interval. However, the algorithms severely overestimated τ^2 when the model contained cubic terms. Especially MetaForest overestimated τ^2 substantially when $\beta = 0.8$ and the estimated model is cubic: $\Delta\tau^2_{MF} = 2.92 \pm 2.37$. The other algorithms also had a $\Delta\tau^2 > 1$ in these conditions, but the results were not as severe. Interestingly, the Pema algorithms even outperformed the RMA algorithm in these conditions.

The marginal effects of β on $\Delta\tau^2$ are shown in image 5B. MetaForest was affected most by the increase in β, but in general performed better than the Pema algorithms when $\beta < 0.8$. The RMA algorithm performed best overall.

The marginal effect of α on $\Delta\tau^2$ was rather minimal, although there was a slight decrease in $\Delta\tau^2$ as α increased. However, the decrease is more explicit when the interaction of α with the estimated model is added. Image 7 shows this interaction. The algorithms were rather unaffected by α for the linear model, and a small decrease in $\Delta\tau^2$ as α increased can be seen in the exponential model. When the two-way interaction model was estimated however, the algorithms benefitted as α increased, while for the cubic model, $\Delta\tau^2$ first increased as α increased from 0 to 2, but decreased as α increased from 2 to 10. RMA performed best, followed by MetaForest. The Pema algorithms performed similarly, but worst.

The effect of the true τ^2 on $\Delta\tau^2$ was rather unnoticeable for the RMA and MetaForest algorithms. The tendency for the Pema algorithms on the other hand, was to overestimate τ^2 more as the true τ^2 increased. Image 5C shows the marginal relationship.

The effect of the number of moderators on $\Delta\tau^2$ was not that large either. A small increase in $\Delta\tau^2$ can be seen in the RMA and MetaForest algorithm as the number of moderators increased which was not found for the Pema algorithms. However, A small note is that MetaForest did substantially increase in $\Delta\tau^2$ as more moderators were added and the estimated model is cubic. Image 8 shows the interaction between the number of moderators and the estimated model.

κ only had a substantial effect for MetaForest; the $\Delta\tau^2$ decreased quite rapidly if κ increased, especially when the cubic model was estimated. For the other algorithms, decreasing κ had little to no effect on correctly estimating the residual heterogeneity. Image 9 shows the interaction of κ with the estimated model.

Finally, the average number of observations in the studies did not have a substantial effect on $\Delta\tau^2$. Image 5D shows the marginal relationship.

Variable selection

To determine the extent to which the algorithms could perform variable selection correctly, the proportion true positives [TP] and true negatives [TN] were calculated. The TP and TN reflect how well the algorithms accredit importance to relevant moderators and discredit importance to irrelevant moderators respectively. It should be noted that TP could only take on values 1 or 0 per simulated iteration, because in all models where $\beta > 0$, only one moderator was simulated to be relevant. As β increased, the already relevant moderator increased in relevance, rather than spreading the relevance over the other moderators. TN had a bigger range and could take on values dependent on how many moderators were taken up in the model. E.g. when $\beta = 0$ and $n_{mods} = n$, $n + 1$ different proportions were possible, $\frac{0}{n}$ up until $\frac{n}{n}$.

There were no differences in variables selected by Highest Density Intervals or Confidence Intervals for both Lasso and Horshoe and so for both algorithms it did not matter which interval type was analyzed. It was found that MetaForest had the highest proportion true positives: $TP_{Mf} = 0.98$, closely followed by RMA: $TP_{RMA} = 0.96$. Horshoe performed slightly better than Lasso; $TP_{Hs} = 0.91$; $TP_{Lasso} = 0.89$. As for TN, it was found that the pema algorithms performed best: TN_{Hs} and $TN_{Lasso} = 0.93$, followed by RMA: $TN_{RMA} = 0.89$. MetaForest performed worst by a large margin: $TN_{Mf} = 0.50$. The Mean Absolute Deviation for all algorithms on both TP and TN was 0, except for MetaForests performance on TN, where the Mean Absolute Deviation was 0.44.

Perfomance on TP and TN were very high for all algorithms, with all mean proportions, except MetaForests performance on TN, exceeding .89. This implies that MetaForest had issues excluding irrelevant moderators from the models. Plots were inspected to determine the effects of the design factors on the proportions. While inspecting the plots, there were found to be little marginal effects of the design factors on TN, while TP was more affected.

Firstly, κ only had a positive effect on TP. As κ increased, TP also increased. MetaForest had the highest TP, followed by RMA and, lastly, Horsehoe and Lasso. The increase in TP for higher values of κ was steeper for the Pema algorithms, however. There was also an interaction of κ with the estimated model shown in image 10. During estimation of the linear and two-way interaction model, the relationship of κ looked relatively linear. At $\kappa = 20$ the TP was relatively low for the algorithms, compared to the cubic model where TP starts at .98 and converges to 1 as κ increased. This latter relationship was also found for the exponential model, although the TP at $\kappa = 20$ was lower.

$\bar{n}$ had a positive and roughly linear relationship with TP for all algorithms. MetaForest performed best, followed by RMA, while Horseshoe and Lasso performed worst. Image 11A shows the relationship.

β had an interaction effect with the model estimated on TN. Only during estimation of the two-way interaction model, TN decreased as β increased, For the other models, TN remained stable. This could be because the interaction model was only fitted when there were 3,4 or 7 moderators, while for the other models, only 2,3 or 6 moderators were used. Image 12 shows the relationship. The effect of β on TP was positive; as β increased, TP increased too.

The true τ^2 had a negative effect on TP, while α seemed to have little effect on both TP and TN. Image 11B shows the marginal relationship of τ^2 on TP.

Finally, there was an effect of number of moderators on TN, but only for the two-way interaction model. The TN increased as the number of moderators did. Image 13 shows the interaction. This relationship was reversed for TP and was found during estimation of all models, i.e. TP decreased as the number of moderators increased. Image 14 shows the relationship.

Discussion

References