

1 Select relevant moderators in meta-regression using Bayesian penalization

2 Caspar J. Van Lissa^{1,2} & Sara van Erp¹

3 ¹ Utrecht University, dept. Methodology & Statistics

4 ² Open Science Community Utrecht

5 Author Note

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

8 Enter author note here.

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

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

Abstract

Todo

Keywords: meta-analysis, machine learning, bayesian, lasso, regularized

Word count: 5356

Select relevant moderators in meta-regression using Bayesian penalization

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 be 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 (Hedges & Vevea, 1998). 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 $i = 1, 2, \dots, k$) are given by:

$$T_i = \Theta + \epsilon_i \quad (1)$$

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

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:

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

→

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:

$$T_i = \Theta + \zeta_i + \epsilon_i \quad (4)$$

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

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

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 $x_{1...p}$ represent the moderators, and $\beta_{1...p}$ the regression coefficients:

$$T_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \zeta_i + \epsilon_i \quad (7)$$

$$(8)$$

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 $\lambda \rightarrow \infty$, 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 the LASSO penalty is but one example of a shrinkage penalty; other penalties exist, with some unique properties.

Bayesian estimation.

Simulation study

The present study set out to validate the BRMA algorithm using a simulation study. As a benchmark for comparison, we used restricted maximum likelihood meta-regression, which is the standard in the field. We evaluated the algorithms' predictive performance in new data, and their ability to recover population parameters. Our research questions are whether BRMA offers a performance advantage over RMA in terms of any of these indicators; under what conditions BRMA does not offer an advantage, and which prior (horseshoe versus LASSO) is to be preferred.

Performance indicators

Predictive performance reflects how well the algorithm is able to predict data not used to estimate the model parameters, in other words, it indicates the generalizability of the model. To compute it, for each iteration of the simulation both a training dataset and a testing dataset are generated. The model is estimated on the training data, which has a varying number of cases according to the simulation conditions. Predictive performance is

then operationalized as the explained variance in the testing data, R_{test}^2 . The testing data has 100 cases in all simulation conditions. The R_{test}^2 reflects the fraction of variance in the testing data explained by the model, relative to the variance explained by the mean of the training data. For a predictive performance measure, it is necessary to use the mean of the training data, as the mean of the testing is a descriptive statistic of that sample. The resulting metric R_{test}^2 is expressed by the following equation:

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

With k being the number of studies in the testing dataset, \hat{y}_{i-test} being the predicted effect size for study i , and \bar{y}_{train} being the mean of the training dataset.

The algorithms' ability to perform variable selection was evaluated by 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 mean deviation of the estimate from the population value, and the variance is given by the variance of this deviation.

Design factors

To examine performance in a range of realistic meta-analysis scenarios, several design factors were manipulated: The number of studies in the training data $k \in (20, 40, 100)$, the average within-study sample size $\bar{n} \in (40, 80, 160)$, the population effect size of relevant moderators $\beta \in (0, .2, .5, .8)$, the number of moderators $p \in (2, 3, 6)$, and residual heterogeneity $\tau^2 \in (.01, .04, .1)$. According to a review of 705 published psychological

meta-analyses (Van Erp et al., 2017), these values of τ^2 fall within the range observed in practice. Note that both BRMA and RMA assume linear effects. To test the robustness of the algorithms to violations of this assumption, true effect sizes were simulated using 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 $\epsilon_i \sim N(0, \tau^2)$. The algorithms further assume normality of residuals. To examine robustness of the algorithms to violations of this assumption, 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. The design factors combined to produce 1944 unique conditions. For all simulation conditions, 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

The RMA algorithm failed to converge for 10 replications, all of which were characterized by low number of cases ($k \leq 40$) and high effect sizes $\beta \geq .5$. These cases were omitted from further analysis.

Predictive performance

Within data sets, the BRMA with a horseshoe prior had the highest predictive performance 50% of the time, followed by RMA, 37%, and finally BRMA with a LASSO prior, 13%. Results indicated that the overall R_{test}^2 was highest for BRMA with a Horseshoe prior and lowest for RMA, see 1. This difference was driven in part by the fact that explained variance was somewhat higher for the BRMA models when the true effect was not null, and by the fact that RMA had larger negative explained variance when the true effect was equal to zero (i.e., there was no effect to detect).

The effect of the design factors on R_{test}^2 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 2 reports the effect size η^2 of simulation conditions on R_{test}^2 .

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 2. 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 1; 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 noted previously, predictive performance of RMA was most negative (negative explained variance) when the effect size was zero. 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). Similarly, the HS algorithm furthermore had the consistently highest predictive performance regardless of number of cases in the training data, and thus also increased less when the number of training cases increased (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, the sensitivity to true positives P and specificity to true negatives N were calculated. Only simulation conditions with $\beta > 0$ were used, such that the effect of the first moderator was always positive in the population and could be used to calculate P , and the effect of the second moderator was always zero in the population and could be used to calculate N .

As the regularized algorithms shrink all coefficients towards zero, it is unsurprising that sensitivity was highest for the un-regularized algorithm 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 regularized algorithms, $N_{HS} = 0.98$, $N_{LASSO} = 0.97$, $N_{RMA} = 0.94$. Overall accuracy reflects the trade off between sensitivity and specificity. As the baserate of true positives and true negatives is equal in this simulation, overall accuracy is simply given by $Acc = (P + N)/2$. Results showed that 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 3, Figure 2) and specificity (Table 4, Figure 3). 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 . Across all design factors, RMA had the highest sensitivity, followed by HS and then LASSO.

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 . Across all design factors, HS had the highest specificity, followed by LASSO and then RMA. 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 the mean and variance of the difference between \hat{b} and β across simulation conditions, respectively.

Across all simulation conditions, HS had the lowest bias for 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 LASSO estimates of τ^2 had the lowest variance, $V_{LASSO} = 1.47$, followed by HS, $V_{HS} = 1.50$, and then RMA, $B_{RMA} = 1.71$.

The effect of the design factors on the bias in τ^2 was evaluated using ANOVAs. Table 5 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.

For the estimated regression coefficient, HS had the greatest (negative) bias across simulation conditions, $B_{HS} = -0.07$, followed by LASSO, $B_{LASSO} = -0.06$, and then

RMA, $B_{RMA} = -0.01$. Note that all algorithms - including RMA - provided, on average, negatively biased estimates. Across simulation conditions, HS had the lowest variance, $V_{HS} = 0.32$, followed by LASSO, $B_{LASSO} = 0.34$, and then RMA, $B_{RMA} = 0.38$.

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 $\hat{b} - \beta$. The skewness of moderator variables had the largest effect on bias in estimated β for all algorithms. Note, however, that this is likely due to the fact that the data simulated with a cubic model are analyzed with a linear model, and thus,

was the estimated model. This was mainly because the algorithms overestimated τ^2 most when the model contained cubic terms. No differences between algorithms in the effect of design factors were observed.

Discussion

This simulation study validated the performance of two versions of the new BRMA algorithm, relative to state-of-the-art meta-regression (RMA). Our analyses examined the algorithms' predictive performance, which is a measure of generalizability, their ability to perform variable selection, and their ability to recover population parameters. Our research questions were whether BRMA offers a performance advantage over RMA in terms of any of these indicators; under what conditions BRMA does not offer an advantage, 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, RMA produced overfit models; in other words, its models generalized poorly to new data. The predictive performance of the BRMA algorithms also suffered less than that of RMA in the presence of more irrelevant moderators. 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). This provides strong evidence that BRMA with a horseshoe prior is generally preferable when the goal is to obtain findings that generalize to new data.

With regard to variable selection, results indicated that the penalized BRMA algorithms had lower sensitivity: they were less able to select relevant moderators than the un-penalized RMA algorithm. Conversely, the BRMA algorithms had better specificity: they were better able to reject irrelevant moderators than RMA. These results are unsurprising because the BRMA algorithms shrink all regression coefficients towards zero. This diminishes their ability to detect true effects and aids their ability to reject irrelevant moderators. 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 whether sensitivity or specificity is more important in the context of their research. When meta-analyzing a heterogeneous body of literature, with many between-study differences that could be coded as moderators, BRMA may be preferred due to its greater ability to retain only relevant moderators. Conversely, when meta-analyzing a highly curated body of literature with a small number of theoretically relevant moderators, un-penalized 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. Note that all algorithms provided, on average, negatively biased estimates. The variance of the estimates followed the opposite pattern. This unsurprising result illustrates the bias-variance trade-off in penalized regression. The greater predictive performance of the

BRMA algorithms is a direct consequence of this trade-off.

We further observed that BRMA with a horseshoe prior had the lowest bias when estimating residual heterogeneity. The BRMA algorithms also had lower variance than RMA when estimating residual heterogeneity. 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 un-penalized model. Together, these results suggest that BRMA has superior predictive performance and specificity, and provides relatively unbiased estimates of residual heterogeneity, relative to RMA.

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 et al., 2017). Another strength is that we made the BRMA algorithms available in a FAIR (Findable, Accessible, Interoperable and Reusable) format by publishing an R package on the “Comprehensive R Archive Network.” 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, we have made all simulation data and analysis code for the present study available online. This code also enables researchers to conduct Monte Carlo power analyses for applied research. Another limitation is that the present study did

not examine the effect of multicollinear predictors. Regularizing estimators typically have an advantage over OLS regression in the presence of multicollinearity; future research ought to examine whether this also applies to BRMA.

A final limitation is that the present study did not examine the effect of dependent data (e.g., multiple effect sizes per study). In principle, the BRMA algorithm can accommodate dependent data by means of three-level multilevel analysis. To our knowledge, there are no theoretical reasons 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.

Another direction for future research is the specification of different priors, aside from the horseshoe and LASSO priors that were examined in this study. To facilitate such research, we provide a generalized BRMA function which is not compiled, and can be fully customized with user-specified priors. The downside of this flexible function is that it is not compiled, and requires the user to set up a compilation toolchain. Compiling the function thus requires some technological sophistication and is more computationally costly.

Although the use of Bayesian estimation has several advantages, one major downside is that Bayesian models are not directly comparable with frequentist models. Another 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

BRMA aims to address the challenge that arises when meta-analysing heterogeneous bodies of literature, with few studies relative to the number of moderators. BRMA can be used to identify relevant moderators when it is not known beforehand which moderators are responsible for between-studies differences in observed effect sizes. To facilitate

adoption of this method in applied research, we have published the function `brma()` in the R package `pema`. Here, we offer several recommendations for its use. The first recommendation precedes analysis, and relates to the design of the meta-analysis. When the search for moderators is exploratory, researchers ought to be inclusive, but focus on moderators that are expected to be relevant, including theoretically relevant moderators, as well as moderators pertaining to the sample, methods, instruments, study quality, and publication type. In our experience, many applied researchers code such study characteristics anyway, but omit them from their analyses for lack of statistical power. Moderators can be continuous or categorical, in which case they should be dummy-coded. Missing data must be accounted for. The best way to do so is by retrieving the missing information, by contacting authors or comparing different publications on the same data. If missing data remains, users can either use a single imputation method (for example, a non-parametric imputation method like `missForest`), or manually aggregate the results across multiple imputations. The effect sizes and their variances must be computed using suitable methods; note that many such methods are available in the R package `metafor` (Viechtbauer et al., 2010). With regard to data analysis, we recommend the use of a horseshoe prior by default, because it demonstrated the best predictive performance and most attractive trade-off between sensitivity and specificity in our simulations.

When reporting results, researchers should substantiate their decision to explore heterogeneity on both subjective and objective grounds. The former can be achieved by simply ascertaining that the body of literature to be meta-analyzed appears to be heterogeneous; the same rationale commonly used to support the use of random-effects meta-analysis (Higgins et al., 2009). The latter can be accomplished by conducting a random-effects meta-analysis without any moderators, and reporting the estimated τ^2 . Note that significant heterogeneity does not constitute sufficient grounds, for deciding to explore ignore heterogeneity, for two reasons: Firstly, because data-driven decisions render any analysis (partly) exploratory, and increase the risk of results that generalize poorly

(i.e., are overfit). The second reason is that tests for heterogeneity are often underpowered when the number of studies is low, and overpowered when it is high, thus limiting their usefulness (see Higgins & Thompson, 2002). As when conducting RMA meta-analysis, researchers should report both the estimated effect of moderators and residual heterogeneity. Regression coefficients can be interpreted as usual, but it is recommended that researchers acknowledge that they are biased towards zero. If all moderators are centered, the model intercept can be interpreted as the overall effect size at average levels of the moderators. Note that, as BRMA is a Bayesian method, credible intervals or highest posterior density intervals should be used for inference, instead of p-values. The null hypothesis is rejected if such intervals exclude zero. As both types of intervals performed identically in the present study, we suggest using credible intervals, which are computationally less expensive. **SARA: Wel of niet nog OLS RMA doen met enkel de significante predictoren van de BRMA?**

Finally, with regards to publication, we highly recommend sharing the data and syntax for the meta-analysis publicly; for example, by making the entire paper reproducible using the Workflow for Reproducible Code in Science (WORCS; REF). Transparency allows readers and reviewers to verify that methods were correctly applied, and try alternative analyses. Particularly when using a new method like BRMA, this transparency is likely to inspire confidence in the results. Secondly, the results of a meta-analysis can be used to obtain predictions for the expected effect size of a new study on the same topic, given specific design characteristics. This prediction 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.

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, suggesting that BRMA is suitable as a small sample solution. This is an appealing quality, 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 keeping overall variable selection accuracy approximately constant 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 may be further facilitated by the availability of the user-friendly R package `pema`.

References

Table 1

Mean and SD of predictive R^2 for BRMA with a horseshoe (HS) and LASSO prior, and for RMA, for models with a true effect ($ES \neq 0$) 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]
$ES \neq 0$	-0.01	[-0.04, -0.00]	-0.01	[-0.02, 0.00]	-0.10	[-0.40, -0.01]

Table 2

Effect size of design factors on predictive R^2 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 3

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 4

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 5

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 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

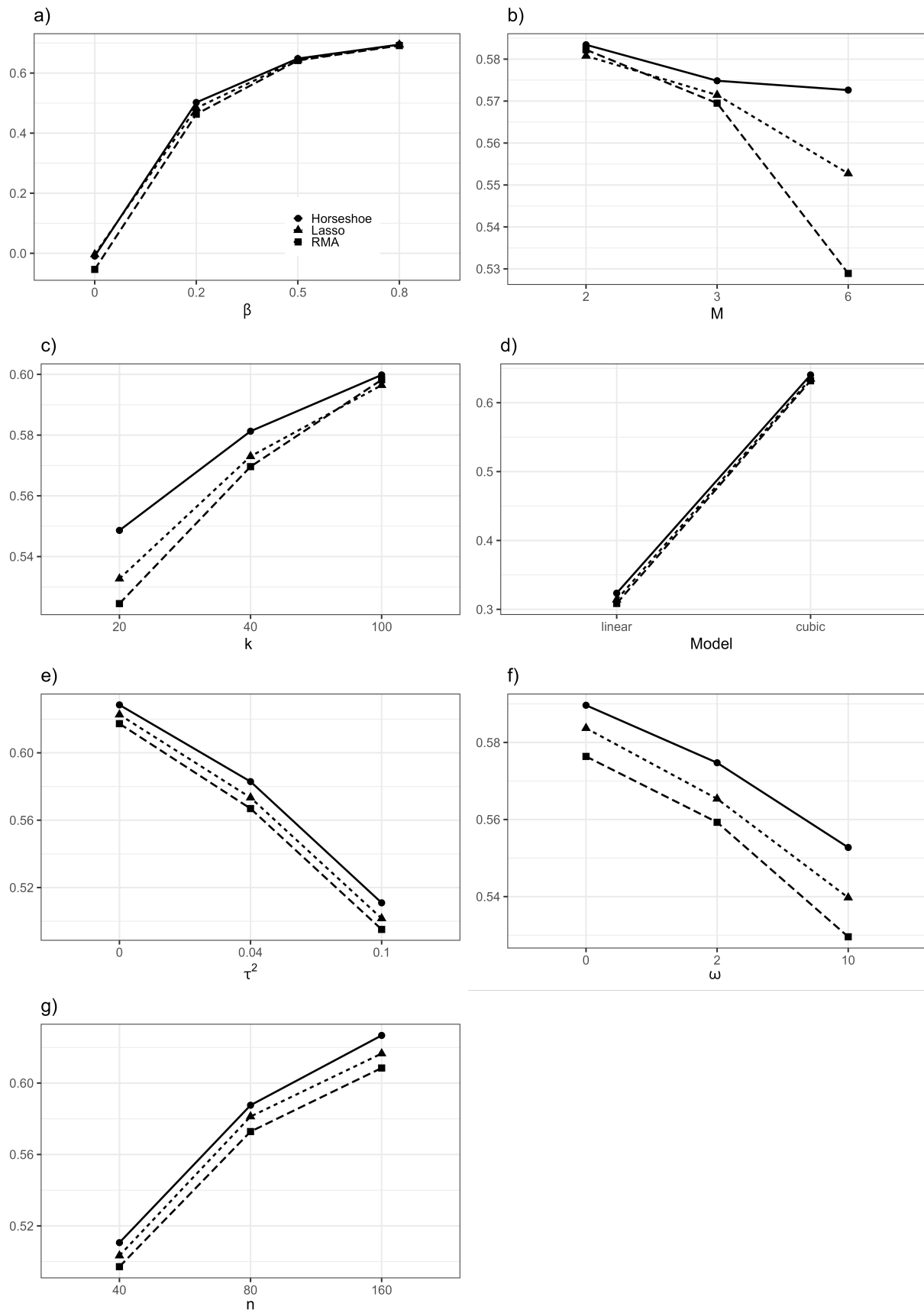


Figure 1. Predictive R2 for BRMA with horseshoe (HS) and LASSO prior, and RMA. Plots are sorted by largest performance difference between BRMA and RMA.

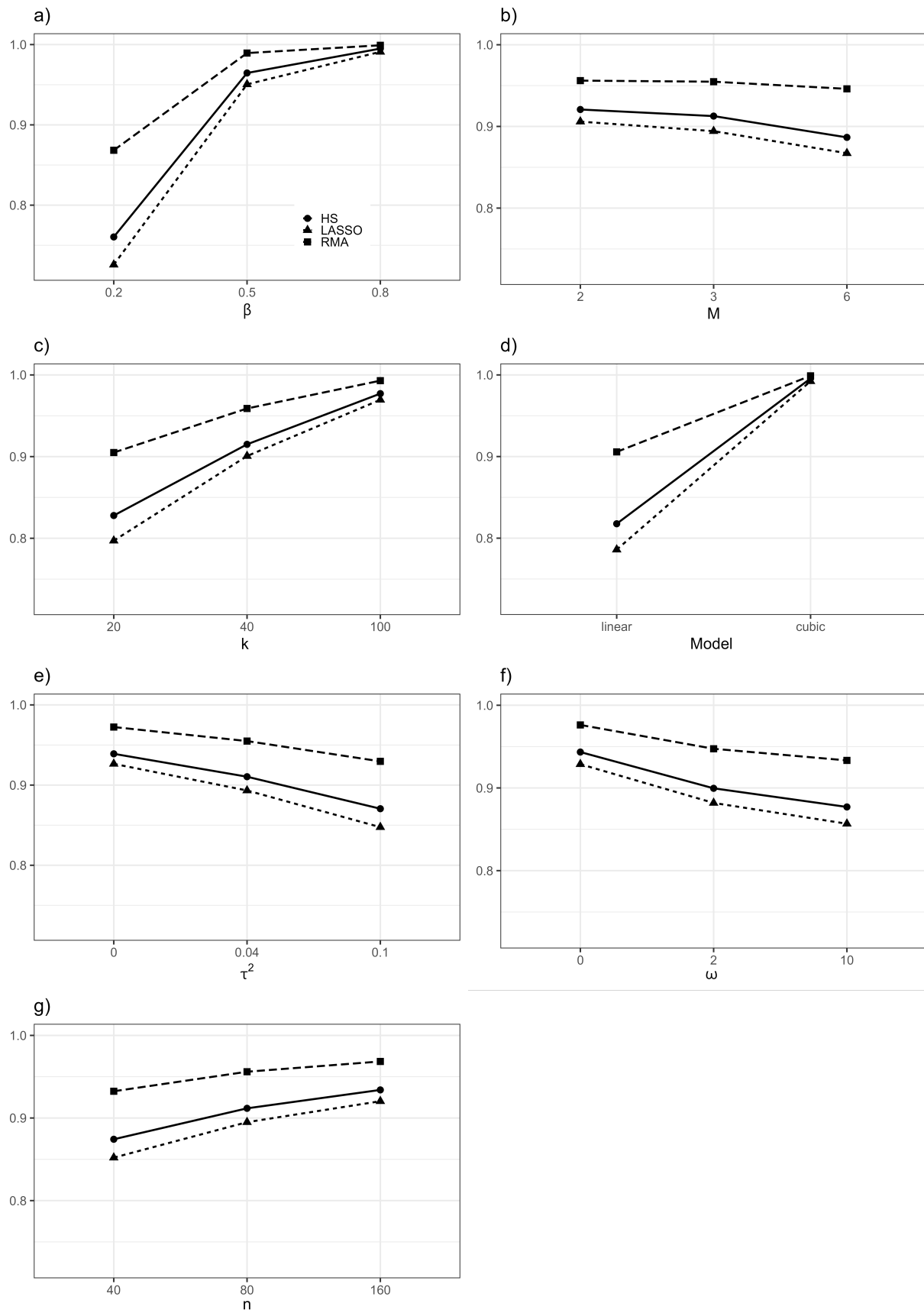


Figure 2. Sensitivity by design factors for the HS (circle, solid line), LASSO(triangle, dotted line) and RMA (square, dashed line) algorithms.

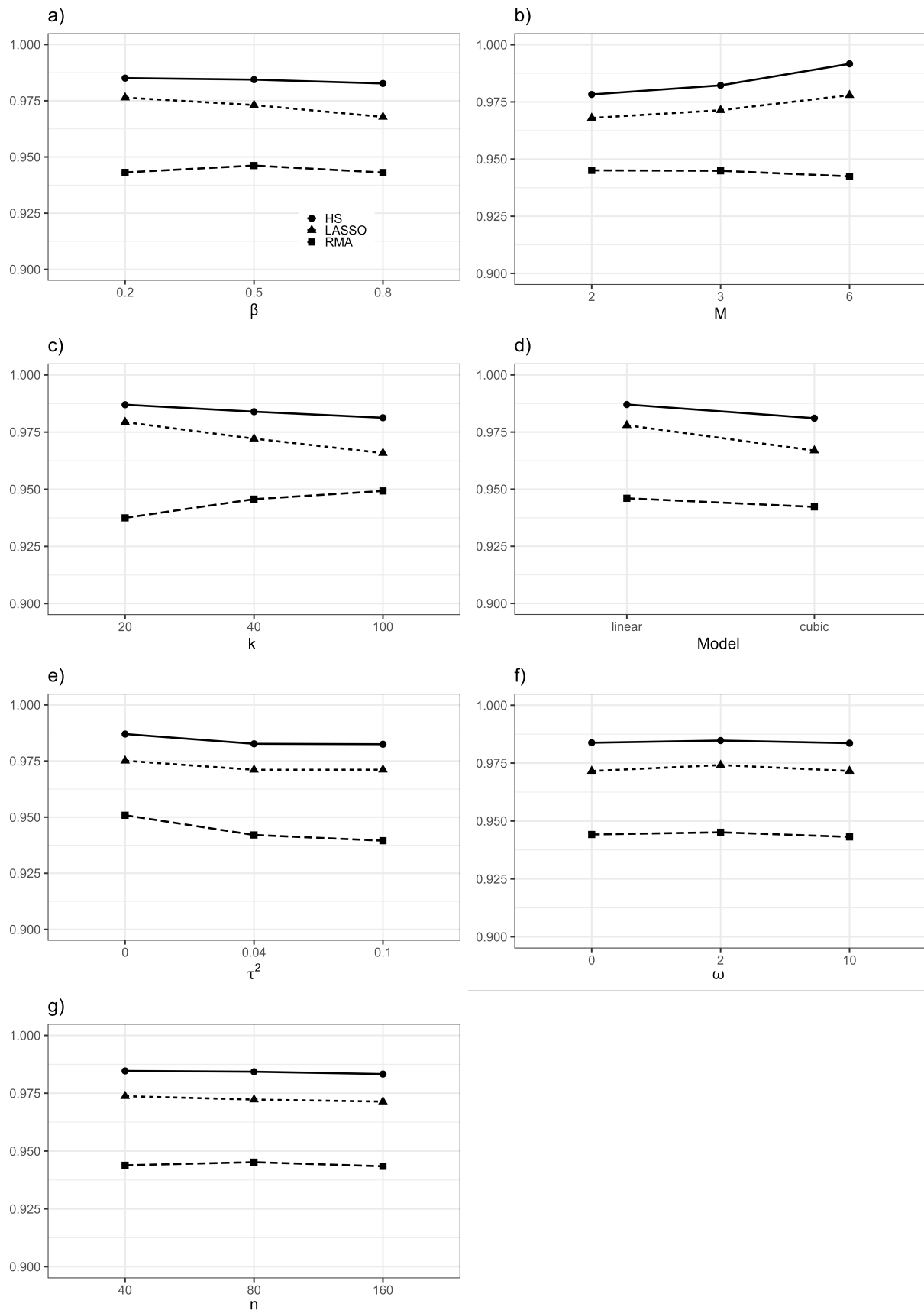


Figure 3. Specificity by design factors for the HS (circle, solid line), LASSO (triangle, dotted line) and RMA (square, dashed line) algorithms.