INTRODUCTION

Meta-analyses often have small sample sizes and moderator analyses are typically exploratory, thereby risking overfitting the data.  Thus, there is a need of a regularization method to curtail overfitting. LASSO (L1-norm regularization) can fulfill this role, since it has an advantage in terms of feature selection. The goal of this project is to implement L1-norm regularization in the weighted meta-regression, developing an new estimator for penalized meta-regression. Meta-regression is usually estimated using either (restricted) maximum likelihood, weighted least squares (DerSimonian and Laird), or even Bayesian estimation. REML (restricted maximum likelihood) method provides estimators that are less biased and hence it will be used in this project. The existing literature does not provide any selection method of both fixed and random effects in mixed-effects meta-regression models. Hence, the problem arises in the simultaneous estimation of regression coefficients, β, and the amount of residual heterogeneity, τ2.

META-ANALYSIS, HETEROGENEITY AND META-REGRESSION

A **systematic review** is a high-level overview of primary research on a focused question that identifies, selects, synthesizes and appraises all high quality research evidence relevant to that question. It uses explicit, systematic methods that are selected with a view to minimizing bias, hence providing more reliable findings from which decisions can be made. **Meta‐analysis**, often an integral part of a systematic review, utilizes statistical techniques to pool data on a common end‐point from studies included in the review.

Meta-analysis is typically a two-stage process. At the first stage, studies fulfilling certain inclusion criteria are obtained and an **effect size** estimate is extracted from each study. An effect size is a statistic of interest that measures the magnitude of a phenomenon. Typical examples of effect sizes are the risk ratio, the mean difference and the correlation between two variables. At the second stage, the study-specific effect sizes are synthesized in a weighted average to produce a summary effect. There are two approaches for synthesizing studies; the **fixed-effect** (FE) and the **random-effects** (RE) **model**. These two models represent fundamentally different assumptions regarding the underlying data.

A fixed-effect model assumes that there is only one source of variation in the studies and that any differences in the observed effect sizes are due to the sampling error. Thus, the fixed-effect model takes into account only within-study variance. A random-effects model allows for the true effect size to differ between studies, and assumes that the differences in the observed effect sizes are attributed not only to the random error, but also to the variation in true treatment effects (**heterogeneity**). Hence, random-effects model takes into account both within- and between-study variances.

Heterogeneity is a measure of the consistency of the results between trials in a meta-analysis. A large degree of heterogeneity occurs when there are large differences between trials in the measured efficacy of an intervention. Meta-analysts often examine to what extent the heterogeneity in the effect sizes can be accounted for, based on various study characteristics (**moderators**). The process of examining the relationship between study characteristics and the effect sizes is typically called a **moderator analysis**. There are two major forms of moderator analysis in meta-analysis: the subgrouping of studies and the meta-regression models.

**Meta-regression** is a more widely used technique that aims to relate the size of effect to one or more characteristics of the studies involved. As multiple regression is used to assess the relationship between subject-level covariates and an outcome, meta-regression in meta-analysis is used to assess the relationship between study-level covariates and effect size. Specifically, the effect size estimates are used as the dependent variable, and the moderators as the independent variables in meta-regression models. Also, a random effect is typically included in such models to account for any **residual heterogeneity** that is not accounted for by the moderators included in the model. Since the predictors included in the model are usually added as fixed effects, this approach then leads to a **mixed-effects meta-regression model**. From a statistical perspective, weighted regression analysis is recommended because it controls for correlations between moderators and does not require moderators to be nested.

MIXED EFFECTS META-REGRESSION MODEL

In a meta-analysis with k independent studies, let **y** denote a *k x 1* vector of independent effect sizes {yi} that represents the results of the studies and **X** a *k x (p + 1)* design matrix of full column rank with p predictor variables, representing some differential characteristics in the studies. Since the predictors are included as fixed effects in the model, assuming a random-effects model for the effect sizes leads to a mixed-effects meta-regression model, which can be expressed by the formula:

**y = Xβ + u + e**, (1)

where **β** is a *(p + 1) x 1* vector containing the regression coefficients {βo, β1, …, βp}, **u** is a *k x 1* vector of independent between-studies errors {ui} with distribution N(0, τ2) and **e** is a *k x 1* vector of independent within-study errors {ei}, each with distribution N(0, vi). While vi is the within-study variance (or sampling error) for the ith study, τ2 represents the residual heterogeneity (or between-studies) variance, that is the remaining variability in the true effect sizes not accounted for after adding one or more predictors to the model.

Note that the mixed-effects model presented in equation (1) is actually an extension of the random-effects model and that the latter can be formulated if **X** is defined as a *k x1* vector of ones. In this case we would have a model without predictors, where **β** is a scalar containing the hypermean (mean of the population effects) and **u** is normally distributed with mean 0 and variance τ2, the latter denoting the total heterogeneity in the true effects. If, moreover, the error term **u** were suppressed from equation (1), then the model would become a fixed-effect model (which is equivalent to setting τ2 = 0 or assuming that the sampling error is the only source of variability). Also, under normality assumptions regarding **u** and **e**, equation (1) imply that **y ~ N(Xβ, τ2I + V)**, where **I** is the identity matrix and **V** is a diagonal with elements vi.

The regression coefficients {βo, β1, …, βp} can be estimated using the weighted least squares formula:

**= (X’ŴX)-1X’Ŵy**, (2)

where **Ŵ** is a *k x k* diagonal matrix with the inverse variances of the effect sizes as elements, which is wi = 1/(vi + 2) for mixed-effects models. 2 represents an estimate of the unknown value τ2. Note that an adequate estimate of both the within-study variance for each study, vi and the residual between-studies variance, τ2, is needed for the estimation of the regression coefficients. For commonly used effect size metrics (e.g. standardized mean differences, correlation coefficients, odds ratios, risk ratios), approximately unbiased estimators are available for vi and the usual practice in meta-analysis is to substitute those estimates and treat them as known values.

ESTIMATION ALGORITHM

Fitting a mixed-effects meta-regression model, it is necessary to estimate not only the model coefficients, but also the amount of residual heterogeneity, τ2, in the effect sizes. Accurate estimation of the residual heterogeneity is a ﬁrst step in this process, as the reduction of residual heterogeneity compared to meta-analysis heterogeneity can be used to interpret whether the study-speciﬁc covariates may or may not explain heterogeneity. Several methods to estimate the amount of residual heterogeneity have been already proposed, 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. The methods are divided into two main groups: closed‐form (or non‐iterative) methods and iterative methods. Closed-form methods provide a parameter estimator in a predetermined number of steps, whereas iterative methods converge to a solution when a speciﬁc criterion is met (however, some iterative methods do not always produce a result because of failure to converge).

The **DerSimonian and Laird** (DL) **method** is the most frequently used approach as it is a non-iterative method that is simple to implement. Many software routines have DL as the default method to estimate the between-study variance. The method assumes that the individual study true effects are distributed with a between-study variance around an “overall” true effect, but makes no assumptions about the form of the distribution of either the within- or between-study effects. Thus, its default use has often been challenged in the sense that DL may underestimate the true between-study variance, potentially producing overly narrow conﬁdence intervals for the mean effect, especially when the between-study variance is large. Likelihood methods are considered as an alternative and effective approach to account for the uncertainty in estimating the heterogeneity by several authors.

**Maximum Likelihood** (ML) **method** is asymptotically efficient but requires an iterative solution. Firstly, it makes an additional assumption that both the within- and between-study effects have Normal distributions, and then the log-likelihood function is solved iteratively to produce an estimate of the between-study variance. However, the method does not always converge while in some cases, the between-study variance estimate is negative and set to zero producing a negative bias. The negative bias of the ML estimate of the between-study variance is a consequence of failing to properly account for the loss of degrees of freedom due to the estimation of the fixed effects.

**Restricted Maximum Likelihood** (REML) **method** is a popular iterative method that takes account of the number of the fixed-effects estimated parameters, losing one degree of freedom for each. This is achieved by applying ML to the least-squares residuals, which are independent of the fixed effects, as the effect of the fixed variables is removed. Hence, REML method estimates parameters that maximize the likelihood of the error distribution while imposing restrictions to avoid over-fitting.

Taking everything into consideration, it is possible to estimate the amount of residual heterogeneity using a closed-form solution (DL method), but an iterative Restricted Maximum Likelihood (REML) estimator is less biased and hence more preferable, especially for continuous outcomes. Once the model has been fitted, the individual model coefficients can be examined to determine the extent to which the moderators are related to the effect sizes.

All the estimators, provided from these methods, can be succinctly expressed after defining the matrix:

**P = W – WX(X’WX)-1X’W**, (3)

where **W** is a diagonal weight matrix whose elements can change from one method to another. For example, for the Hedges method, **W** is defined as the identity matrix. When weights are included, non-iterative estimators (e.g., DerSimonian-Laird estimator) make use of the inverse of the within-study sampling variances, while for the iterative estimators (e.g., maximum likelihood estimator) **W** contains the inverse variances plus an estimate of the amount of residual heterogeneity.

The underlying logic for all these methods is to estimate the residual heterogeneity based on the difference or ratio between some estimate of the total variability among the population effect sizes not accounted for by the explanatory variables included in the model and the amount of variability expected from random sampling error alone.

In particular, the total variability not accounted for by the explanatory variables is expressed as a quadratic form of the effect size estimates and (a function of) the **P** matrix. For example, with the elements of the diagonal weight matrix **W** set equal to wi = 1 / vi, we obtain the residual heterogeneity statistic **QE = y'Py**, which is simply equal to the residual sums of squares under weighted least squares estimation. On the other hand, the amount of sampling variability in the effect size estimates is given by the vi values, which we can collect in the diagonal matrix **V**. The value of the quadratic form, relative to (some function of) **V** and/or the degrees of freedom of the model under assessment (i.e., df = k - p - I), then provides an estimate of the residual amount of heterogeneity.

BIAS-VARIANCE TRADE OFF

All learning algorithms use a mathematical approach that contains an “error” term which can be further split into two components: reducible and irreducible error. The irreducible error is an inherent uncertainty associated with the model and is associated with a natural variability in a system. This cannot be reduces and nothing can be done about it. On the other hand, reducible error can be and should be minimized further to maximize accuracy.

In supervised learning algorithms, this reducible error can be further decomposed into “error due to bias” and “error due to variance”. The goal of the learning algorithm is to simultaneously reduce bias and variance in order to obtain a model that is the most feasible. However, achieving this is not easy in reality, since there is always a a trade-oﬀ between these values.

**Bias** is the difference between the average prediction of our model and the correct value which we are trying to predict. The error due to bias appears when a model is “too simple” for the data that it is trying to predict. This usually occurs because one way to speed up the learning phase is using a simple objective function that is easy to compute. However, it also results in making too many simplifying assumptions and hence ending up with a high bias model (underfitting).

**Variance** is the variability of model prediction for a given data point or a value which tells us spread of our data. The error due to variance arises when a model is “too complex” for the data it is trying to fit. When the learning algorithm is very influenced by the specifics of the training data, every time we use a different sub-set we get a quite different prediction model. And the more sensitive the algorithm, the more different the model, and the more variable the prediction of the same input point (overfitting).

The more complex the model is, the more data points it will capture and the lower the bias will be. However, complexity will make the model “move” more to capture the data points and hence its variance will be larger. In other words, bias has a negative first-order derivative in response to model complexity while variance has a positive slope.

In general, finding an optimal bias-variance tradeoff is a difficult task, but acceptable solutions can be found, e.g., by means of cross-validation or regularization methods. **Cross-validation** is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. **Regularization** is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.

REGULARIZATION

**Ordinary Least Squares** (OLS) method is one of the most commonly used prediction techniques, in which the OLS estimates are obtained by minimizing the residual squared error. However, there are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is prediction accuracy: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is interpretation. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects. Hence, it is obvious that OLS estimation does not perform well with variable selection.

An alternative strategy is the **Regularized Regression** methodsallowing to create a linear regression model that is penalized, for having too many variables in the model, by adding a constraint in the equation. This is also known as shrinkage regression or penalized regression methods. By adding this penalty, the regularization technique constrains/shrinks the coefficient estimates towards zero, avoiding the risk of overfitting. Note that, the shrinkage requires the selection of a tuning parameter (lambda), which controls the amount of shrinkage.

The most common types of regularization techniques used to address over-fitting and feature selection are L1 Regularization and L2 Regularization. A regression model that uses L1 Regularization technique is called **LASSO Regression**,whereas a model that uses L2 Regularization technique is called **Ridge Regression**. Both of these techniques, work by penalizing the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. The key difference is in how they assign penalty to the coefficients.

Ridge regression shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero. In Ridge Regression, the shrinkage of the coefficients is achieved by penalizing the regression model with the penalty term called L2-norm, which is the sum of the squared coefficients. In the case of LASSO (Least Absolute Shrinkage and Selection Operator) Regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero. LASSO shrinks the regression coefficients toward zero by penalizing the regression model with the penalty term called L1-norm, which is the sum of the absolute coefficients.

The main advantage of LASSO over Ridge method is that it enforces the irrelevant coefficients to be exactly zero, thus it removes some features altogether. So LASSO regression not only helps in reducing over-fitting, but it also works well for feature selection.

LARS ALGORITHM

The computation of the LASSO solutions is a quadratic programming problem and can be tackled by standard numerical analysis algorithms. However, the **Least Angle Regression** (LAR) procedure is a better approach. LAR can be viewed as a kind of “democratic” version of forward stepwise regression and is intimately connected with the LASSO, providing an extremely efficient algorithm for computing the entire LASSO path.

Supposing that *i = 1, …, k* is the number of studies and *j = 1, …, p* is the number of predictors, the LAR algorithm has the following form:

1. Start with r = y, 1, 2,..., p =0. Assume Xj standardized
2. Find predictor Xj most correlated with r
3. Increase βj in the direction of sign(corr(r, Xj)) until some other competitor Xk has as much correlation with current residual as does Xj
4. Move (j, k) in the joint least squares direction for (Xj, Xk) until some other competitor Xl has as much correlation with the current residual
5. Continue in this way until all predictors have been entered. Stop when corr(r, Xj)=0 ∀ j, i.e. OLS solution

The LARS algorithm includes LAR as well as LASSO or forward stepwise as implemented by least-angle methods. A simple modification of the LAR algorithm that gives the entire LASSO path is as follows:

1. Start with LAR
2. If a coeﬃcient crosses zero, stop. Drop that predictor, recompute the best direction and continue

The LARS algorithm is extremely efficient, requiring the same order of computation as that of a single least squares fit using the *p* predictors. Though LAR always takes *p* steps to get to the full least squares estimates, the LASSO path can have more than *p* steps, since predictors can drop out. If *p k – 1,* we have the solution after min(k - 1, p) steps.

SELECTION PROCEDURE AND PROPOSED METHOD

As mentioned above, Restricted Maximum Likelihood (REML) method provides estimates with the less bias comparing to Weighted Least Squares (WLS) method. Thus, we will make use of REML method to estimate the unknown parameters, adding the Least absolute shrinkage and selection operator (LASSO) penalty in the likelihood function to curtail the risk of data overfitting. After that, we will use the Least Angle Regression (LARS) algorithm to get all the LASSO solutions.

For **y ~ N(Xβ, τ2I + V)**, the restricted log-likelihood function with the LASSO penalty of the parameter vector (β, τ2) is given by:

***llREML* = - ½kln(2π) + ½ln|X’X| - ½ln|τ2I+V| - ½ln|X’WX| - ½y’Py + λj|**, (4)

where X is a *k x (p + 1)* matrix containing a vector of ones and the values of the p moderator variables, I is the *k x k* identity matrix, V is the diagonal matrix with elements vi, W is the diagonal weight matrix with elements wi = 1 / (vi + τ2), P is a matrix given by equation (3) and y is a *k x 1* vector of observations. τ2 and β are the unknown parameters of the between-study variance and regression coefficients, respectively. βj is the jth element of the vector β, which is also unknown. The tuning parameter that controls the amount of regularization is denoted as λ (λ 0).

The penalized REML estimator of (β, τ2) is that set of values that maximizes *llREML* under the constraint: τ2  0. Firstly, if we exclude the adding penalty, the *llREML* is a function of two unknown parameters that we want to estimate. Since the equation (2) corresponds to the REML estimator of β for a given value of τ2, it can be substituted in the *llREML* function. Thus, the formula provided by this substitution is the equation (4). Following the same logic, equation (4) can be simplified in another equation of just one unknown parameter, τ2. Given the equation (2), we can get the unknown **j|** after some calculations. Hence, the final *llREML* will have the following form:

***llREML* = - ½kln(2π) + ½ln|X’X| - ½ln|τ2I+V| - ½ln|X’WX| - ½y’Py - λij’ŴiXij|-1|Xij’Ŵiyi|]**, (5)

where the only unknown parameter is τ2. Therefore, the problem simplifies to finding that value of τ2 that maximizes equation (5). We will denote this value by 2.

Since there is no closed-form solution for obtaining 2, iterative computations are required. There are several procedures that can be used for this purpose, but Fischer scoring algorithm is the most commonly implemented algorithm for maximizing the likelihood function in the R package “metafor”. The Fischer scoring algorithm is known to be robust to poor starting values and it usually converges quickly. After getting the value of 2, the can be obtained through the LARS algorithm.

FISCHER SCORING ALGORITHM

The maximization of the likelihood function can be performed using several techniques, such as the Fischer scoring algorithm, the Newton-Raphson method, the Simplex method or the Expectation-Maximization (EM) algorithm. For the Fischer scoring algorithm, the starting value of 2 is set equal to the value obtained with any of the other noniterative estimators, like the DerSimonian and Laird:

**2DL = [y’Py – (k – p – 1)] / tr(P)**, (6)

where P is defined in the equation (3) and W is a diagonal matrix with elements given by wi = 1 / vi (note that y’Py = i(yi - 0 - 1Xi1 - … - pXip)2, where j, *j = 1, …, p*, are the estimates obtained from equation (3)).

The initial estimate is then adjusted based on a factor Δ (the inverse Fischer information of τ2 multiplied by the first derivative of the profiled log-likelihood with respect to τ2), yielding a new estimate 2new. This process continues until convergence and can be expressed by:

**2new = 2current + Δ**, (7)

where 2current is the current estimate of τ2. For REML estimation, the adjustment factor can be shown to be equal to:

**ΔREML = [y’PPy – tr(P)] / tr(PP)**, (8)

With P defined in equation (3) and W is a diagonal matrix with elements given by wi = 1 / (vi + 2current). Therefore, after each step, we first update W, then P and finally we can compute ΔREML to obtain 2new. The iterative process terminates when ΔREML is smaller than some preset threshold (e.g., ΔREML 10-5).

An additional complication arises, because equation (7) may yield a negative value of 2new. This problem can be easily avoided by using step halving. For this, we check on each iteration whether 2current + ΔREML  0 and if this is the case, we continue to multiply ΔREML by 1/2 (i.e., first by ½, then by ¼, then by 1/8 and so on) until ΔREML becomes small enough, such that 2new stays nonnegative. This ensures that the final value obtained for 2 is also nonnegative.

LIMITATIONS

Even though the proposed method is encouraging, there are some limitations in the estimation process of the unknown parameters (β, τ2) that should be addressed.

A major difficulty of this method is the computational cost, since REML is a method of estimation for variance parameters, which requires an iterative scheme. This means that a lot of calculations are required to get the desirable solution and hence more time is needed. However, after comparing other iterative and non-iterative methods, it was noticed that the use of REML estimation is the most preferable.

Furthermore, through the estimation procedure there are different methods and algorithms combined. This could cause confusion to the accurate and simultaneous estimation of (β, τ2), since there are several parameters that we should be aware of.

CONCLUSION

The proposed method for the simultaneous estimation of (β, τ2) is the implementation of LASSO (Least Absolute Shrinkage and Selection Operator) regularization in weighted meta-regression and specifically, in REML (Restricted Maximum Likelihood) estimation. After that, the Fischer scoring algorithm is suggested to obtain the unknown parameter 2 and then, the unknown parameter can be derived from the LARS (Least Angle Regression) algorithm. This way, data overfitting is curtailed and in the same time the estimates of the unknown parameters (β, τ2) are obtained.

Besides some limitations of this method, Penalized Restricted Maximum Likelihood estimation is a rigorous method, in which the adjustment of overfitting is directly built into the model development. It is a promising method that combines different procedures to get the estimates of the unknown parameters and that could have future expansions for the prediction models.