Regularization in mediation models: A Monte Carlo simulation comparing different Regularization penalties in multiple mediation models

Submitted by Arjunvir Ghumman

Supervised by Ji Yeh Choi

A thesis submitted to the Faculty of Graduate Studies in partial fulfilment of the requirements for the degree of Master of Arts

Quantitative Methods in Psychology

York University

Toronto, Ontario

August 2022

Abstract

The two fundamental goals in statistical learning are establishing prediction accuracy and discovering the correct set of predictors to ensure model specificity. Although the field of variable selection has made significant strides over the past decades, these methods are yet to be fully adapted to mediation models. Regularization methods that utilize the l_1 penalty such as the Lasso and adaptive Lasso incorporate a small amount of controlled bias into the ordinary least squares estimates to help improve the generalizability of the estimates by significantly reducing their variance across samples. Additionally, the Lasso can perform variable selection and help achieve model selection consistency or sparsistency. Recent literature has proposed methods that have introduced regularization to mediation models. These include regularized structural equation modelling or RegSEM. The current research compares the performance of various regularization penalties such as the Lasso, adaptive Lasso, MCP and SCAD in the context of mediation models. No single regularization penalty performed optimally across all simulation conditions. Additionally, we observed disproportionate selection rates for the Lasso and SCAD penalty with alternating mediators which was indicative of disproportionate shrinkage of the *a* and *b* pathways. However, the absolute bias induced in the *a* and *b* pathways was equivalent across all samples for each penalty term. This highlights the perils of shrinking individual regression pathways instead of indirect effects as a whole. Overall, the choice of the type of regularization penalty implemented depends on the particularities of the research question.

Keywords: *l*1-penalty; Regularization; Bias-Variance Trade-off; Sparsity; Variable selection; mediation

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

When fitting linear models, the goal is to assign coefficients to each predictor variable. In some instances, the number of variables *P* is large, and we may seek a smaller subset of variables to enhance interpretability. If interpretability of models is of paramount importance, we should strive to achieve simple structure; this relates to the field of sparsity. A sparse model consists of only a small number of non-zero parameters. Hastie, Tibshirani and Wainwright (2016) highlighted several advantages of choosing a sparse model over a dense model such as the interpretation of the fitted model, improved prediction accuracy and computational convenience (due to the convexity of the solutions which allows for scalable algorithms that allow us to handle dense problems). The authors also delineated one further advantage, which arises from mathematical literature and is termed the "*bet on sparsity principle*" and states that *"we should use a procedure that does well in sparse problems, since no procedure does well in dense problems"* (Hastie et al., 2016, p. 2).

In certain cases when the number of hypothesized predictors exceeds the number of cases *n*, the data is high dimensional (if the number of predictors is larger than the sample size) and the regression coefficients are not uniquely defined as the matrix of the predictors cannot have linearly independent columns and hence is not invertible. Consequently, there are an infinite set of solutions that allow the objective function to be zero, and these solutions almost surely overfit the data as well. Therefore, there is a need to constrain or regularize the estimation process. When dealing with multiple predictors, researchers may be interested in exploring a subset of predictors which explains most of the observed variability in the outcome variable. Regularization methods such as the Lasso allow us to select the relevant variables by constraining or regularizing the coefficients of interest towards zero and sets some of them exactly to zero, effectively filtering them out of the final model. In statistics, a large body of work has been devoted to the topic of variable selection. Eminent statistician

and researcher, Bradley Efron, went as far as to identify this as the most important area of research in the field of statistics (Hesterberg, Choi, Meier, & Fraley, 2008).

With a recent surge in popularity, mediation models have become increasingly ubiquitous in psychological research. Various authors (Pieters, 2017; Fairchild & Mcdaniel, 2017) have described mediation as vital to theory development and an indispensable tool to evaluate potential causal effects using intervening variables (mediators). Mediation is often described as a third variable effect that elucidates how a predictor variable relates to an outcome variable. There has been an influx of methodological research on mediation analysis (Fairchild, MacKinnon, Taborga & Taylor, 2009; Kelley & Preacher, 2012; MacKinnon & Dwyer, 1993; MacKinnon, 2012) since Baron and Kenny's (1986) seminal article was published.

To study the impact of a predictor on an outcome variable through another variable, we rely on mediation analysis. This framework helps us understand the causal process of the relationship between the predictor and the outcome variable. When dealing with multiple mediators, Structural Equation Modelling (SEM) is considered the preferred framework (e.g., Preacher & Hayes, 2008). However, SEM requires fitting regression-type models with mediators as predictors; hence, the estimates are prone to overfitting when the number of mediators become close to the sample size. One possible solution is to fit the SEM model with a subset of the predictors using some selection criteria. Another solution is to apply regularization methods (e.g., penalized regression) that shrink the regression coefficients towards zero or in some instances, exactly to zero, resulting in a sparse model with the true subset of mediators related to the outcome. In recent literature, several methods have been proposed to deal with such cases, and many such methods have been conveniently placed under the umbrella term "Exploratory Meditation Analysis" or EMA (Preacher & Hayes, 2008; Serang, Jacobucci, Brimjall & Grimm, 2017). More specifically, Serang and colleagues

extended the use of regularization methods to identify a sparse subset of mediators in the context of a mediation model. The proposed study aims to compare the performance of various regularization methods available to researchers in the context of mediation analysis using a simulation study design.

The remainder of this research paper is organized as follows. The rest of chapter 1 overviews the multiple mediation model along with approaches to testing mediation and effect sizes for mediation. Later we introduce regularization methods and the *l*1 penalty (specifically, the Lasso) along with generalizations of the Lasso penalty that helps achieve sparsity. Sparsity in the context of mediation refers to a small number of mediators with nonzero mediation effects between a treatment and an outcome; hence, only the mediators that explain the relationship between the predictor and outcome variables are included in the final model. Finally, to conclude chapter 1, we describe various methods for regularization in mediation models while delineating each method's advantages and limitations and highlighting the proposed research's goals. Chapter 2 describes the simulation design and conditions. Chapter 3 discusses the results of the Monte Carlo simulation design. Chapter 4 expounds on the implications of the current research, further examining the limitations and directions for future research, and Chapter 5 reports the overarching conclusions we can draw from this paper.

1.1 Mediation Models

The mediation process can be conceptualized as a third-variable effect that explains how or why two or more variables relate to each other. The theory of mediation analysis implies that a causal process exists and seeks to examine the impact of a mediating variable, which is hypothesized to transmit the influence of a predictor onto an outcome variable. In social sciences, mediation analysis is typically conducted to examine the underlying mechanisms by which a predictor relates to the outcome variable.

Before we examine the theory behind mediation analysis, it is helpful to briefly revisit the historical foundations of the mediation models. The influential work of Judd and Kenny (1981) outlined several requisite conditions for mediation to take place. Specifically, the explanatory variable must impact the outcome variable, the explanatory variable must impact the mediator, the mediator must impact the outcome variable, and the explanatory variable must not impact the outcome when the mediator is controlled. Baron and Kenny (1986) reiterated the conditions mentioned above. Additionally, they expounded that the relationship between the explanatory variable and the outcome must simply be reduced in magnitude rather than set to zero when we control for the mediator, as posited by Judd and Kenny. Further work by Shrout and Bolger (2002) further relaxed the condition requiring a significant effect of the independent variable on an outcome for mediation to occur. These conditions have been widely referred to as the causal steps.

We begin by describing the mediation model in its simplest form, when a single mediator, *M*, is postulated to mediate the effect of the predictor variable, *X*, to the outcome variable, *Y* (MacKinnon, 2012). Alwin and Hauser (1975) helped illuminate how the estimation of mediation effects is well-rooted in path analysis, which is a special case of structural equation modelling (SEM) that allows us to extend regression analysis by providing us with a framework for evaluating whether a set of data fits a causal model of interest. We begin by describing the mediation model in its simplest form, when a single mediator, M, is postulated to mediate the effect of the predictor variable, X, to the outcome variable, Y (MacKinnon, 2012). The mediation model posits that the three variables relate to each other via a causal chain of effect, such that the predictor variable in the model impacts the mediator and the mediator subsequently impacts the outcome variable which occurs last in the chain. Hence, in this way, the predictor variable acts on the outcome variable through the mediator.

For any path analysis, a corresponding path diagram illustrates the hypothesized causal relations in the model by denoting directionality (represented by single-headed arrows). Every path diagram has a corresponding system of equations that delineates the statistical model. The mediation model can be described using a set of three regression equations (MacKinnon & Dwyer, 1993),

$$Y = i_1 + cX + e_1$$

$$Y = i_2 + c'X + bM + e_2$$

$$M = i_3 + aX + e_3 .$$

In the above equations, the *c* coefficient refers to the total effect of the predictor on the outcome variable, whereas, *c'* describes the direct effect of the predictor on the outcome variable controlling for the mediator. The *a* and *b* regression weights define the effect of the predictor on the mediator and the effect of the mediator on the outcome variable (controlling for the predictor variable), respectively. The i_1, i_2 and i_3 and e_1, e_2 and e_3 terms represent the intercept terms and the corresponding residuals for the three regression equations, respectively. Figure 1 illustrates this relationship using a path diagram.

The general assumptions underlying the mediation model include those associated with regression analysis as the mediation model can be conceptualized in terms of multiple regression equations. These include the correct specification of the model (no omitted variables), no measurement bias in the predictors, no correlated residuals, correct causal ordering of the variables and no presence of reverse causality effects (Fairchild & McDaniel, 2017). Much of the work in social science research involves examining causal behaviour and relationships. Hence, the authors outlined the necessary assumptions required for the indirect effect to be considered causal. Estimating the causal structure among the variables is often not possible if the sample is not randomly assigned to the levels of the mediator. In regression analysis, it is often a requirement that the predictor is randomized to avoid correlating the

predictor with the residual variance of the mediator or the outcome. An additional assumption requires the a priori belief that a causal effect exists for the mediator-outcome path. Several solutions have been proposed to deal with causal inference problems. VanderWeele (2010) proposed several methods for conducting sensitivity analyses via marginal structural models, which can be used to assess the robustness of indirect effects when dealing with the omitted variable problem. These methods allow us to explain further why a causal ordering of a mediation model may no longer apply after modelling confounder variables. Finally, the authors also described how incorporating design features into a mediation model may eliminate alternative explanations of effects by promoting heterogeneity.

Testing Mediation Models

There are three common approaches to testing mediation models in the literature, joint significance testing, difference in coefficients and product of coefficients. The joint significance testing approach for mediation evaluates the statistical significance of the regression weights *a* and *b* in the above regression equations. If both regression coefficients are significant, mediation has occurred. One of the major limitations of the joint significant approach is that it fails to provide a single point estimate or standard error (SE) for the mediated effect (MacKinnon et al., 2002).

On the other hand, the difference in coefficients approach compares the regression coefficients *c* and *c*'. The goal is to determine whether there is a significant reduction in the regression coefficients when the mediator is controlled for. If the direct effect of the *X* on *Y* reduces significantly when we control for the mediator *M*, then mediation is said to have occurred assuming all model assumptions are satisfied. MacKinnon and colleagues (2002) suggested that conceptualizing the mediation effect in this way may lead to the conflation of mediators and confounders due to a lack of directionality. Although the difference in coefficients approach provides us with a single point estimate and standard error estimate for

the mediated effect, it fails to provide us with the necessary framework to estimate regression coefficients and test the significance of their difference when dealing with two or more mediators (MacKinnon et al., 2002).

The third and final approach is the product of coefficients. The product of the coefficients approach is widely regarded as the gold standard, especially when implementing mediation models in the SEM framework. This approach was derived from path analysis models and utilizes path tracing rules to estimate the indirect effect of the explanatory variable on the outcome variable through the mediator. This approach involves multiplying the regression weight *a* with the regression weight *b*. Fairchild and McDaniel (2017) elucidated that the multiplication of the two regression weights accounts for the impact of the explanatory variable on the mediator and the impact of the mediator on the outcome variable (controlling for the explanatory variable). Furthermore, the product of the coefficients approach can easily be extended to more complex models involving multiple mediators. Various authors have recommended using the third and final approach, product of coefficients, for testing mediation models (Preacher & Hayes, 2008; MacKinnon et al, 2002; Fairchild & McDaniel, 2017).

We can test for the significance of the mediator using Sobel's test (Sobel, 1982). Sobel's test entails dividing the product of coefficients, ab, by a delta estimate of the SE and comparing it to a standard normal distribution to test for significance (*Ho*: $ab = 0$). Sobel's method to extract a Confidence Interval (CI) was initially applied with larger samples in mind as the central limit theorem denotes that samples drawn from a non-normal population will approximate a normal distribution as the sample size grows larger. However, it is unreasonable to expect the indirect effects to follow a normal sampling distribution when dealing with smaller samples. Although Sobel's test provides us with SE estimates, their use is not recommended as they impose distributional assumptions and restrict us from

formulating asymmetric CI's to test significance for the mediator (Fairchild & McDaniel, 2017). Additionally, Shrout and Bolger (2002) recommended using modern resampling methods such as bootstrapping to estimate the SE's of the indirect effects, as these methods do not impose any assumptions of normality on the sampling distribution of the indirect effect. Various other authors (MacKinnon et al., 2002; Preacher & Hayes, 2008) have also recommended resampling procedures such as bootstrapping over Sobel's test to maintain control over Type 1 error rates. Bootstrapping requires resampling of the dataset with replacement and estimating the indirect effect in the resampled dataset. This process is typically repeated *n* times (with *n* preferably being greater than 1000) to obtain *n* different estimates of the indirect effect (Preacher & Hayes, 2008). The empirical bootstrap makes no assumptions regarding the distribution of the indirect effect and only requires it to be a good approximation of the underlying population distribution. Although it is impossible to estimate the true confidence interval of a statistic, the bootstrap method allows us to asymptotically estimate the true confidence interval as the sample size increases (Preacher & Hayes, 2008).

Effect Sizes

Traditional statistical significance testing informs us that an effect exists based on normal theory. However, as previously stated, the underlying assumption that the indirect effects are distributed normally is erroneous. Furthermore, it provides no information about the size or magnitude of the effect, which aids us in understanding the practical or clinical significance of the effect.

An effect size can be described as the quantification of some phenomenon of interest that allows us to address a specific research question (Kelley & Preacher, 2012). The authors proposed several properties that would be required for a statistic to serve as an effective measure of an effect size. First, the statistic should be on an interpretable scale; it is often good practice to standardize the effect size measures to ensure they are comparable across

studies. Second, it should be possible to estimate confidence intervals based on the sampling distribution of the statistic in question. Third, the statistic needs to be unbiased, consistent and efficient. Fourth, the effect size should be a monotonic function of the effect it quantifies, and finally, the population effect size should be independent of the sample size.

In regression analysis, several statistics can serve as effective measures of effect size, including standardized mean differences, correlation coefficients and proportion of variance measures. Standardized regression coefficients on their own serve as valid measures of effect size. However, Lachowicz, Preacher and Kelley (2018) argued that these measures do not translate well into mediation analysis as the effect of interest is the indirect effect is a product of two regression coefficients and has an unknown underlying distribution.

Historically several effect size measures have been proposed. The following passage provides a brief yet comprehensive review of the two effect sizes that are relevant to the current study along with their limitations.

Preacher and Kelley (2012) proposed κ^2 as a measure which the authors defined as the ratio of the observed indirect effect with the maximum possible indirect effect. Wen and Fang (2015) argued that κ^2 is not an appropriate effect size measure for mediation models due to its lack of the property of rank preservation since κ^2 is not a strictly monotonic function and an increase in indirect effect does not necessarily lead to an increase in the κ^2 estimate. Additionally, κ^2 is only appropriate for simple mediation models with only a single mediator and has not been generalized to multiple mediation models. The authors demonstrated that these issues are due to the improper estimation of the maximum possible value of the indirect effect. Proportion of variance measures were proposed due to their intuitive appeal and ease of implementation. However, simulation studies have demonstrated that these methods have severe limitations (Preacher & Kelley, 2011; Wen & Fang 2015).

Lachowicz, Preacher and Kelley (2018) proposed a novel measure upsilon (*υ)* for quantifying the explained variance in mediation analysis. This new measure accounts for the variance in the outcome jointly by the predictor and mediator whilst correcting for spurious correlations that may be induced by the ordering of the variables. The *υ* measure has key advantages over other measures as it outperforms other R^2 measures when dealing with suppression as it maintains the monotonicity of the indirect effect in the population when suppression is evident.

Multiple Mediation Model

Preacher and Hayes (2008) described the multiple mediation model as a simple extension of the single mediator model where additional pathways are included to represent multiple mediators. These pathways increase the number of ways by which the explanatory variable can influence the outcome variable. The multiple mediation model with *j* mediators can be expressed using the following three regression equations. Figure 2 illustrates the multiple mediation model with *J* mediators

$$Y = i_1 + cX + e_1$$

$$Y = i_2 + c'X + \mathbf{\Sigma}^{J}_{j=1} b_j M_j + e_2$$

$$M_j = i_{j+2} + a_j X + e_{j+2},.$$

where *X* refers to the predictor variable, *Y* refers to the outcome variable and M_j refers to the *j*th mediator where $j = 1,2,3,..., J$. The regression coefficient *c* represents the total effect of *X* on *Y,* which remains unchanged from the single mediator model, whereas *c*' is the direct effect of *X* on *Y* adjusted for the presence of the mediators. The i_1, i_2 and i_{J+2} and e_1, e_2 and e_{j+2} terms represent the intercept terms and the corresponding residual terms. Since the model includes *J* mediators, instead of estimating a single indirect effect, we estimate *J* separate indirect effects, each representing the predictor's effect on the outcome controlling for other mediators. Each of these indirect effects is simply referred to as a specific indirect effect. The

specific indirect effects are determined by multiplying the two specific paths (a_jb_j) linking X to Y through the mediator M_j. The total indirect effect of X on Y is estimated by summing up the specific indirect effects ($\mathbf{\Sigma}^{J}_{j=1}a_jb_j$).

Preacher and Hayes (2008) explicated several advantages of specifying a multiple mediation model over several single mediation models. Firstly, if the total indirect effect is significant, we can conclude that the set of J mediators facilitates the relationship between a and b. Secondly, it allows us to determine the degree to which specific mediators mediate the relationship between a and b, conditional on the presence of other mediators (assuming the mediators are correlated to each other). Thirdly, it reduces the likelihood of parameter bias due to omitted variables as specified by Judd and Kenny (1981). And finally, the presence of several mediators in the model allows us to pit competing theories under the same model.

1.2 Regularization

Regularization methods were developed to deal with overfitting issues when fitting models with the goal of striking an optimal balance between fitting the data and finding an optimal sparse solution (Hastie, Tibshirani, & Friedman, 2009). Before we delve deeper into the world of regularization, it is pertinent to briefly introduce the concepts of bias, variance, and overfitting.

Overfitting in Regression Models

In statistical sciences, we strive to test our statistical model against a novel dataset collected under different circumstances to evaluate how well the results generalize. If our model successfully extracts the true population characteristics, it will not only perform well in the sample dataset but also generalize to new datasets. Babyak (2004) described overfitting as the tendency to capitalize on the idiosyncratic information of the sample at hand. Simply put, random noise becomes entangled with the signal, which leads to finding spurious effects when dealing with small to moderate sample sizes. Overfit models are true to their name,

they overperform when fit to the sample data but tend to perform far less admirably when fit to other samples from the same population. The concept of overfitting can be illustrated using a very simple example. If we attempt to estimate the mean from a population, the *SE* of the mean (the average amount the sample estimate will differ from the population parameter) would be much larger for smaller samples compared to larger samples which suggests that the sample estimate of the mean would fluctuate more when the sample size is relatively small. Similarly, for regression models, the regression weights will fluctuate considerably more when dealing with smaller samples. This instability of regression weights increases the likelihood of obtaining an overly optimistic fit. It is important to note that overfitting does not lead to biased estimates in the sample data as the estimates from an overfit model perform very well on the sample data where they were originally fit (McNeish, 2015). They simply fail to generalize to other samples. Hence, overfit models have inflated regression weights, small *SE*'s, large R^2 values and small *p* values. Furthermore, simple structure (parsimony) is not achieved as extraneous predictors may be erroneously included in the model (Babyak, 2004).

Bias-Variance Tradeoff

Assuming that the true regression model is known, the statistical bias of the regression weights is given by the following equations, where $E(\theta')$ refers to the expected value of the sample regression weights, θ' denotes the sample regression weights and θ denotes the population regression weights. Bias can be defined as the expected difference between the expected sample regression weights and the population weights. We can state that the sample is unbiased if the $E(\theta')$ is equal to θ.

$$\text{Bias}(\theta') = E(\theta') - \theta$$

Helwig (2017) defined Mean Squared Error (MSE) as the total expected squared difference between the estimated and true regression weights. MSE is often expressed in the

following form, where MSE stands for the mean squared error, $V(\theta)$ denotes the total variance of the regression weights and $bias^2$ (θ') is the total squared bias.

$$MSE(\theta') = V(\theta') + bias^2\ (\theta')$$

The MSE quantifies both the accuracy (bias) and precision of the regression weights (variance). Ideally, we would prefer to estimate regression weights with low bias and low variance; however, this is not always the case in practice. Unbiased regression weights minimize the prediction error for a single sample but tend to vary considerably more between samples. Conversely, biased estimates may not be ideal for the sample dataset; however, they may vary less across different samples of the population provided that the estimates are not too biased (McNeish, 2015).

Ridge Regression

In linear regression, our goal is to approximate the outcome variable using a linear combination of predictors. We parameterize our model by a vector of regression weights ($\boldsymbol{\beta}_1 \ldots\ldots \boldsymbol{\beta}_p$) where P is the number of predictors. In certain cases, our regression weights may have low bias but high variance (multicollinearity, overfitting); these are some of the few reasons why we might consider implementing ridge regularization. Regularization methods intentionally introduce bias into the sample to reduce variance across other samples from the population. Ridge regression enforces a penalty that constrains the sum of the squared regression coefficients (Hoerl & Kennard, 1970). Using the sum of squared regression coefficients (aka an l_2 penalty) results in shrinkage which is proportional to the size of the regression weight (larger regression weights are shrunk more towards 0 compared to smaller regression weights) (Hesterberg et al., 2008).

$$\boldsymbol{\beta}(\text{ridge}) = \boldsymbol{\beta}_{OLS} \quad \text{subject } to\ ||\boldsymbol{\beta}||^2 \leq t^2$$

$$\boldsymbol{\beta}(\text{ridge}) = \boldsymbol{\beta}_{OLS} + \lambda \boldsymbol{\Sigma}^{P}_{j=1} \boldsymbol{\beta}^2.$$

The above equations denote the ridge regression penalty (l_2) and the ridge regression penalty in Lagrangian form. $\boldsymbol{\beta}_{\text{OLS}}$ refers to the least-squares estimator for the regression weight $\boldsymbol{\beta}$, which is obtained by minimizing squared-error loss. The $\boldsymbol{\Sigma}^{P}_{j=1}||\boldsymbol{\beta}||^2 \leq t^2$ refers to the l_2 norm constraint where t acts as a budget constraint that limits the sum of squared values of the parameter estimates. When the budget constraint is relaxed $\boldsymbol{\beta}$(ridge) is the same as $\boldsymbol{\beta}_{\text{OLS}}$; however, when a budget constraint is applied the parameter estimates are shrunk towards zero. The budget is specified by an external procedure such as the cross-validation. $\hat{\lambda}$ serves as a tuning parameter that directly controls the amount of shrinkage. A larger value of $\hat{\lambda}$ results in more shrinkage towards zero and a smaller value results in minimal shrinkage. If $\hat{\lambda}$ = 0, $\boldsymbol{\beta}$(ridge) is the same as $\boldsymbol{\beta}_{\text{OLS}}$. The $\boldsymbol{\beta}_{\text{OLS}}$ term is typically divided by a factor of $\frac{1}{2}N$ this corresponds to the reparametrization of λ which allows us to compare $\hat{\lambda}$ values for different sample sizes (useful for cross-validation) (Hastie et al., 2016). It is important to note that the ridge regression model does not perform variable selection unless a specific threshold is specified to completely eliminate certain variables from a model.

Lasso (least absolute shrinkage and selection operator)

Another reason why we might consider implementing regularization methods is for the purposes of interpretation. In certain cases we might have a large number of predictors or as in the case of high-dimensional problems the predictors outnumber the cases ($p>n$), the goal is to often identify a smaller subset of these predictors that exhibit the strongest effects. The least absolute shrinkage and selection operator (Lasso) was introduced by Tibshirani (1996) with the goal of reducing overfitting and identifying the correct subset of predictors. Lasso minimizes the absolute value of the sum of the regression coefficients (an l_1 penalty). As previously mentioned, as the penalty term grows larger, the regression coefficients are shrunk to zero. The l_1 penalty allows some regression coefficients to be shrunk to exactly zero which promotes sparsity (a sparse statistical model sets some parameters to 0 resulting in a

sparse solution). This allows the Lasso to perform variable selection in addition to obtaining regularized regression coefficients. The Lasso applies an equal amount of shrinkage across all parameters, which allows certain parameters to be set to exactly 0 unlike ridge regression, which shrinks coefficients proportional to their size (Hastie et al., 2016).

$$\boldsymbol{\beta}(\text{Lasso}) = \boldsymbol{\beta}_{\text{OLS}} \quad \text{subject to } \boldsymbol{\Sigma}^{P}_{\text{j=1}}|\boldsymbol{\beta}| \leq \text{t}$$

$$\boldsymbol{\beta}(\text{Lasso}) = \boldsymbol{\beta}_{\text{OLS}} + \lambda\boldsymbol{\Sigma}^{P}_{\text{j=1}}|\boldsymbol{\beta}|.$$

The above equations represent the Lasso. $\boldsymbol{\Sigma}^{P}_{\text{j=1}}|\boldsymbol{\beta}| \leq \text{t}$ represents the l_1 norm constraint where the bound t acts as a budget constraint similar to the ridge regression. For Lasso, large values of t allow the model to adapt closely to the training data. In contrast, smaller values of t restrict parameters (setting some to zero), leading to a sparse solution. Similarly, for the Lagrangian form, the λ (tuning parameter) controls the amount of shrinkage applied. Before performing Lasso, it is standard procedure to standardize the predictors unless the predictors are measured in the same units (Hastie et al., 2016).

It is important to note that there is a certain correspondence between the constrained problem and the Lagrangian form; this means for each value of t in the range where the constraint (l_1 or l_2) is active, there is a specific value of λ that yields the exact same solution from the Lagrangian form (Hastie et al., 2016).

Cross-Validation

A value of t that is too small prevents the Lasso from differentiating signal and noise, while too large a value can result in overfitting. That's why it is of the utmost importance to strike a balance and find a value of t that provides a sparse solution while minimizing the prediction error (Hastie et al., 2016). The first step of cross-validation involves splitting the dataset into k folds (k is typically equal to 5 or 10), where we typically allot the kth fold as the test set and the remaining k-1 folds as the training set. We later repeat this process a total of k times, with each fold serving as the test set for a range of different values of t (The range is

designated as *n*). We later apply the regularization method to the training data to *n* different values of *t* (we typically begin with a small value of *t*), and we use each fitted model to predict the response in the test set. This allows us to estimate the mean squared error (*MSE*) for each value of *t*. Next, we average the *MSE* across each fold to obtain *n* different mean squared estimates along with some measure of variability. We can plot the *MSE* for a range of *n* values to obtain the *MSE* curve. Finally, we compare the mean squared error for each value of *t* and select the smallest value of *t* or a value of *t* that is within one standard error of the minimum *MSE* value (Hastie et al., 2016).

1.3 Generalizations of the l_1 penalty

Relaxed Lasso: While the bias introduced by Lasso plays a vital role in improving the generalizability of parameter estimates, it's not as efficacious in exploratory settings. Or perhaps researchers might be interested in the variable selection property of the Lasso and wish to obtain unbiased estimates for their sample. After obtaining a sparse solution using the Lasso, we can debias the parameters by applying the least squares estimation process to the subset of predictors with nonzero coefficients in the Lasso. This two-stage process is also known as the relaxed Lasso (Meinshausen, 2007).

SCAD (smoothly clipped absolute deviation) and MCP (minimax concave penalty): There are a few noted limitations of the Lasso: inconsistent selection and high false-positive rates. To overcome these issues, alternative forms of regularization methods have been proposed. SCAD consists of a nonconvex penalty which was one of the earliest and most influential when proposed (Wang, Li, & Tsai, 2007). The SCAD penalty operates similarly to the Lasso as it retains the penalization rate (and bias) of the Lasso for coefficients with a smaller magnitude but continuously relaxes the rate of penalization as the absolute value of the coefficient increases. The idea behind the minimax concave penalty (MCP) is quite similar. As with SCAD, MCP starts out by applying the same rate of penalization as the

Lasso, then relaxes the penalization rate to zero as the magnitude of the coefficient increases. In comparison to SCAD, the MCP relaxes the penalization rate immediately, while with SCAD, the rate is reduced more slowly (Zhang, 2010).

Non-Negative Garrote and the Adaptive Lasso: The non-negative garrote by (Breiman, 1995) shrinks smaller values of the regression coefficients more severely than Lasso, and the opposite for larger values. Hence, for larger values of β, the shrinkage factor will be close to 1 (little to no shrinkage), but if the β estimate is small, it will be shrunk toward zero. Non-Negative Garrote also performs variable selection. Zou (2006) proposed the adaptive Lasso, which uses a weighted penalty to shrink certain estimates more severely compared to larger estimates, similar to the non-negative garrote. The weighted estimates can be obtained using the OLS (Ordinary Least Square) regression coefficients when the number of predictors is less than the sample size. Using the OLS regression coefficients to estimate the weights, results in good recovery properties under certain conditions (Huang, Ma and Zhang 2008). The adaptive Lasso is computationally equivalent to the Lasso. The adaptive Lasso scales each individual parameter by their least-squares estimates. Hence each parameter receives a unique penalty relative to its magnitude, resulting in lower false-positive rates. Additionally, the adaptive Lasso has the Oracle property, which suggests that it always recovers the true model characteristics under certain conditions. Furthermore, the adaptive Lasso retains the oracle property in high-dimensional settings under the right conditions (Zou, 2006). The adaptive Lasso and non-negative garrote attempt to reduce the bias while retaining the variable selection property of the Lasso.

1.4 Regularization in Mediation Models

In social sciences, mediators are generally judiciously selected prior to analysis based on substantive theory or expertise. However, in certain cases, the available degrees of freedom may be limited (due to a large number of mediators or smaller samples) or the

researcher may be engaging in novel research where previous theory is limited and the goal is to identify a subset of variables that could potentially mediate the relationship between the explanatory variable and the outcome variable. Under these conditions, regularization methods may be applied to mediation analysis to achieve sparsity or reduce overfitting to improve the generalizability of the results. MacKinnon (2012) suggested a two-step approach that required conducting mediation analysis (including all possible mediators) in the first step and including those mediators deemed statistically significant in step 1 in the final model. However, the author expressed concern over the large number of tests that will utilize many degrees of freedom and may result in inflated Type 1 error rates and recommended exercising some control over the family-wise error rates. However, Babyak (2004) argued that such methods tend to overfit the data under regression settings.

Jacobucci (2017) extended the regularization methods to Structural Equation Modelling (SEM*)*. They termed the novel method Regularized Structural Equation Modelling (RegSEM*)*. RegSEM penalizes specific parameters in structural equation models, intending to identify the correct subset of predictors and achieve a sparse solution (Jacobucci, Grimm, & McArdle, 2016). RegSEM builds the regularization penalties directly into the estimation process, which lends researchers additional flexibility in testing various models. The expected covariance matrix in RegSEM is estimated using the information from the Reticular Action Model (RAM) matrices (see, Jacobucci, Grimm, & McArdle, 2016) and is later implemented into a maximum likelihood (ML) function. Similar to a regression framework, when lambda is zero, *ML* estimation is performed and as the penalty grows, the selected parameters are shrunk towards zero. RegSEM currently allows users to implement various regularization penalties, namely, Lasso, MCP, SCAD and adaptive Lasso. Furthermore, RegSEM's framework allows us to choose which parameters to penalize. This

provides additional flexibility, which permits penalization of certain parameters or all of them at once based on existing theory.

Serang and colleagues (2017) proposed a novel method named "XMed" which imposed the l_1 penalty on mediation pathways using RegSEM. This method consists of a two-step approach inspired by the relaxed Lasso (Meinshausen, 2007). The first step involved imposing the Lasso penalty, which shrunk small regression paths to zero, identifying a subset of potential mediators. The second step involved debiasing the estimates by refitting the mediation model with the subset of mediators obtained in step 1. This method shrinks certain regression paths to 0, irrespective of the value of their contemporary associated paths. In the first part of the simulation study, the authors compared the variable selection property of the XMed with more traditional *p-value* methods for a multiple mediator model. The five mediator model included in the analysis was split into the noise and effect groups with three mediators (small, medium and large) serving as the effect and two mediators with a population indirect effect of zero serving as the noise. The authors concluded that XMed consistently outperformed traditional *p-value* methods in identifying the correct set of mediators. In the second part of the study, the authors examined the bias induced by XMed and discovered that the relaxed Lasso was particularly effective for debiasing the estimates.

1.5 Current Study

The current study aims to build upon the work of Serang and colleagues (2017) and further optimize the selection of mediators in an exploratory framework. Figure 3 illustrates the simulation design for the study. XMed demonstrated the superiority of the Lasso in identifying the correct subset of mediators compared to more traditional *p-value* based methods. Additionally, other studies have demonstrated that other regularization penalties such as the Adaptive Lasso, SCAD and MCP outperform the Lasso in identifying the correct

subset of predictors in regression settings. Using a series of Monte Carlo simulations, the current study will investigate the following research questions:

1. Which of the four specified regularization penalties leads to optimal selection rates for the signal variables and minimizes the Type I error rates for the noise variables?
2. Does RegSEM shrink the *a* and *b* paths disproportionately, and does that impact selection rates?
3. If the pathways are being shrunk disproportionately, what is the extent of the bias induced?

Based on existing theory, I hypothesized that the selection rates for the simulation design would vary based on the conditions. Lasso, MCP and SCAD would perform similarly when dealing with smaller effect sizes; however, the bias induced in MCP and SCAD would be lower than Lasso for medium to large effects. The adaptive Lasso would penalize each parameter relative to its magnitude, resulting in the lowest false-positive rates. Additionally, the bias induced would vary for other regularization penalties compared to the Lasso as they apply a unique penalty depending on the magnitude of the effect size.

As the sample size increases, the adaptive Lasso is expected to outperform all other methods in selection rates and minimizing the false positive rate due to its oracle property. Furthermore, we would expect both *a* and *b* pathways to be shrunk equally and the selection rates would not vary depending on the pathway. The following section describes the simulation setup and design.

Chapter 2. Method

To investigate the variable selection properties and the bias induced by different regularization penalties in the context of a multiple mediation model, a series of Monte Carlo simulations were performed using R (R Core Team, 2021). The simulation factors and the data generation process were kept consistent with the parent study (Serang et al., 2017) to further examine how well the current study replicates the selection rates for the Lasso penalty. The mediation models were estimated using the maximum likelihood estimator in the lavaan package (Rosseel, 2012) and the regularization penalties were implemented using the RegSEM package (Jacobucci, Grimm, & McArdle, 2016). Mediation effects were carefully chosen for the population model and compared with sample estimates of indirect effects for 1,000 replications of the study design. The two simulation factors were the sample size and the type of regularization penalty implemented. The four regularization penalties were Lasso, adaptive Lasso, SCAD and MCP and each penalty was implemented across a range of different sample sizes (50, 100, 200, 500, 1000 and 2000). In total, there were 24 design rows (6 x 4), and indirect effects were estimated for each design row and compared to the population models. Figure 4 illustrates the simulation design for the current study.

2.1 Study Design

The population model consisted of seven separate mediators, with five mediators serving as signal variables and two serving as noise variables. The inclusion of the noise variables was justified as they allowed us to make in-model comparisons and directly compare the variable selection property of each regularization penalty along with their Type-I error rates. The first three signal variables were chosen to map onto small, medium and large effect sizes as recommended by Lachowicz, Preacher and Kelley (2018). The authors recommended using the standardized regression coefficient benchmarks ($M_S = 0.14$, $M_M = 0.39$, $M_L = 0.59$) proposed by Cohen (1998) for the regression pathways *a* and *b*. Hence the

population indirect effects for the first three signal variables were as follows, small = 0.02, medium = 0.15 and large = 0.35. The other two signal variables were included to examine the relative shrinking of the *a* and *b* pathways. These mediators had different effect sizes mapped onto the *a* and *b* regression pathways. The *a* pathway for the fourth signal variable was mapped onto a small effect size and the *b* pathway was mapped onto a large effect size (M_{SL}). The reverse was true for the fifth and final signal variable where the *a* pathway was mapped onto a large effect size and the *b* pathway was mapped onto a small effect size (M_{LS}). The two noise variables were mapped onto a regression coefficient of zero for both *a* and *b* pathways to denote no population effect (M_N). Finally, the direct effect of the predictor variable *X* on the outcome variable *Y* was mapped to zero. *Figure* 1 illustrates the population model used by Serang and colleagues (2017) and *Figure* 2 contrasts their model with the population model utilized in the current study.

The first part of the study aimed to examine the variable selection properties for each specified regularization penalty. The regularization penalty was varied for each subsequent sample size allowing for direct comparison across a range of sample sizes. The data were generated using a set of regression equations that were previously specified and the residual terms were generated from a standard normal distribution. All variables were scaled and centred prior to analysis as is custom when implementing regularization penalties. All regression pathways were penalized when implementing RegSEM. To determine the optimal value of lambda, 120 lambda values of lambda were selected ranging from 0 to 0.357 increasing in increments of 0.003. The lambda value that resulted in the minimum BIC value was chosen to serve as the penalization rate for the model. The regularization model was optimized using the coordinate descent method with a tolerance level of 10^{-6}. Finally, the model was refit using the optimal chosen value of lambda and mediators with nonzero indirect effects were considered selected. In contrast, mediators that were shrunk to zero were

filtered out of the final model. To determine the effectiveness of each regularization method, the proportion of replications for which each signal variable was correctly selected for each design row was calculated. The Type I error rate was determined by calculating the proportion of replications where the noise variables were incorrectly selected in the final model. The optimal regularization method would be one which correctly identifies all signal variables and filters out the noise variables. Finally, the selection rates were compared for M_{SL} and M_{LS} within each regularization penalty to compare the relative shrinking of *a* and *b* regression paths. Recall that these mediators had opposing effects on the *a* and *b* paths.

For the second part of the study, the absolute bias in both the *a* and *b* pathways was estimated for each regularization method. This induced bias is a trademark of all regularization methods and varies depending on the regularization penalty implemented. The first step in determining the absolute bias involved computing the absolute difference between each estimated pathway and the population pathway for all replications. Later the average of these 1000 replications was used to signify the absolute bias for each pathway. The absolute bias estimates were obtained for the first three signal variables (M_S, M_M, M_L) and each variable had two respective pathways; hence six estimates of absolute bias were obtained for each design row.

Chapter 3. Results

3.1 Model Convergence and Proper Solutions

Model Convergence rates for each design row are displayed in Table 1. Across all conditions and replications, 99.27% of the sample models successfully converged. The design rows denoting the smallest sample size were the only ones that failed to converge. For the highlighted design rows with a sample size of 50, 98.9% of the models converged for the adaptive Lasso, 99% of the models converged for MCP, 99.4% of the models converged for Lasso and 99.6% of the models converged for SCAD. Of these estimated models, 95.61% of the models produced proper solutions. As the sample size increased from 50, we obtained a 100% convergence rate with all models yielding proper solutions. Solutions that did not converge or yield proper solutions were excluded from further analysis.

3.2 Mediator Selection for Signal and Noise Variables

Table 4 displays the mediator selection rates for Lasso and adaptive Lasso, and Table 5 shows the mediator selection rates for SCAD and MCP.

Lasso Penalty

For the first three signal variables with small, medium and large effect sizes (M_S, M_M, M_L), Figure 5 compares the selection rates across sample sizes for the Lasso penalty. The Lasso penalty started correctly identifying M_L 99% with a sample size of 50. However, the Lasso penalty required a sample size of 100 to start correctly identifying M_M 97.9% of the time. As the sample size grew beyond 100, the Lasso penalty began to correctly identify M_M and M_L 100% of the time. The Lasso penalty initially struggled to correctly identify M_S, identifying it 69.5% of the time with a sample size of 200. As the sample size grew to 500, the Lasso penalty began identifying M_S 95.1% of the time. The effect of sample size on mediator selection rates for the three signal variables is apparent, with selection rates increasing as the sample size increased. For the two noise variables, Lasso reported high Type

I error rates, and they increased as the sample size grew larger. The Type I error rates for Lasso ranged from 17% to 19% for a sample of 50 and increased to 27% for a sample size of 2000. Figure 9 displays the Type I error rates for the Lasso penalty. Finally, I compared the final two signal variables with differing regression paths. We observed that the selection rates across the two mediators varied even with reasonably large sample sizes. M_{SL} consistently had lower selection rates compared to M_{LS}, with differences ranging from 13% for a sample size of 50 to 3% for a sample size of 500. The differences grew smaller as the sample size increased, with both mediators achieving a 100% selection rate with a sample size of 1000. Figure 13 displays this difference graphically.

Adaptive Lasso

Figure 6 compares the selection rates of the adaptive Lasso penalty for the small, medium and large effect-sized mediators across our range of sample sizes. Adaptive Lasso struggled to correctly identify M_S and M_M with a sample size of 50 with selection rates of 9.8% and 73.8%, respectively. The adaptive Lasso penalty struggled to identify M_S even with larger sample sizes and required a sample size of 1000 to identify the mediator with a 94.4% selection rate. Adaptive Lasso began consistently identifying M_M and M_L with a sample size of 100 with selection rates of 93.7% and 100%, respectively. Figure 10 displays the Type I error rates for the adaptive Lasso penalty. The Type I error rates for the adaptive Lasso penalty were almost negligible, with an error rate of approximately 2% with a sample size of 50 and less than 1% error rates for each subsequent sample size. Finally, Figure 14 compares the selection rates of M_{SL} and M_{LS}. The selection rates did not display an overarching pattern with differences ranging from 6% in favour of M_{LS} and 3% in favour of M_{SL}.

SCAD

Figure 7 graphically compares the selection rates for M_S, M_M and M_L for the SCAD penalty across all sample sizes. The SCAD penalty displayed unique patterns of selection

rates. When dealing with a sample size of 50, the selection results were similar to those obtained by the Lasso, with the SCAD identifying M_S, M_M and M_L 38.8%, 92.7% and 99.7% of the time, respectively. However, as the sample size grew, the SCAD penalty followed the pattern displayed by the adaptive Lasso. The SCAD penalty required a sample size of 1000 to correctly identify the M_S 92.8% of the time. A similar pattern was observed with noise variables with SCAD reporting high Type I error rates for smaller sample sizes but negligible Type I error rates as the sample size increased. The Type I error rates ranged from 20% to 22% for a sample size of 50 and were less than 1% when the sample size increased to 500. The Type I error rates are displayed in Figure 11. When examining the differences between M_{SL} and M_{LS}, the pattern displayed by the SCAD penalty was remarkably different from the one observed by the Lasso penalty. M_{SL} consistently had higher selection rates compared to M_{LS} with differences ranging from 17% to 4% for the designated range of sample sizes. The differences levelled off for sample sizes greater than 1000. Figure 15 compares the selection rates for the two mediators graphically.

MCP

Figure 8 displays the selection rates for the MCP penalty for the first three mediators. The MCP penalty reported high selection rates starting with a sample size of 50. M_S, M_M and M_L had selection rates of 88.4%, 99.7% and 100% respectively. The MCP correctly identified the three signal mediators in most replications. The MCP required a sample size of 200 to identify M_S 94.2% of the time and optimal selection rates were obtained for all sample sizes greater than 200. Figure 12 depicts the Type I error rates for the MCP penalty graphically. For lower sample sizes, the Type I error rates were very high, ranging from 66% to 85% for sample sizes up to 200. However, similar to the SCAD penalty, the Type I error rates exponentially decreased as the sample size increased, with MCP reporting error rates ranging from 13.7% to 15.6% as we reached a sample of 2000. Finally, the differences in selection

rates for M_{SL} and M_{LS} were examined, similar to other regularization penalties. No meaningful differences were detected between the two mediators as the selection rates were very high even when dealing with smaller samples. The selection rates for M_{SL} and M_{LS} are presented in Figure 16.

Figures 17 to 23 compared the mediator selection property of each regularization penalty across the range of sample sizes. These Figures depict the selection rates for M_S, M_M, M_L, M_N, M_N, M_{SL} and $M_{LS,}$ respectively.

Signal and Noise Variables

The MCP penalty consistently outperformed other regularization penalties across all sample sizes when selecting the correct subset of signal variables. However, the Type I error rates were extremely high, although they decreased exponentially as the sample size grew. For large sample sizes (2000), only the adaptive Lasso penalty identified the correct subset of mediators 100% of the time whilst not selecting the noise variables in any of the 1000 replications. SCAD penalty performed similarly to the Lasso for small to moderate sample sizes; however as the sample size grew larger, the patterns of variable selection displayed by SCAD mimicked the adaptive Lasso penalty.

For the Lasso penalty, M_{LS} reported higher selection rates across all sample sizes compared to $M_{SL,}$ indicating that the *a* pathways were being disproportionately shrunk to zero compared to the *b* pathways. We observed a similar pattern for the SCAD penalty; however in the case of SCAD, M_{SL} reported higher selection rates across all sample sizes compared to M_{LS}. This was in direct contrast to the Lasso penalty as the *b* pathways were being disproportionately shrunk to zero compared to the *a* pathways. We observed no such patterns for the MCP and adaptive Lasso penalties.

3.3 Bias Induced in Signal Mediators

Table 2 and Table 3 depict the absolute bias induced in the *a* and *b* pathways for each regularization penalty, with Table 2 reporting the absolute bias for the Lasso and adaptive Lasso and Table 3 reporting the absolute bias induced in SCAD and MCP.

The overall trend of absolute bias induced due to the Lasso penalty was similar across the three mediators M_S, M_M and M_L. For M_S, the absolute bias induced in the *a* pathways decreased incrementally from 0.067 to 0.018 as the sample size increased from 50 to 2000. Comparatively, the absolute bias induced in the *a* pathways decreased incrementally from 0.062 to 0.017 as we moved from a sample size of 50 to 2000. Similar patterns were observed for M_M as the absolute bias estimates ranged from 0.105 to 0.042 for the *a* pathway and 0.101 to 0.037 for the *b* pathway, with larger samples consistently reporting less bias. Finally, the same pattern persisted for M_L, with bias estimates ranging from 0.152 to 0.095 for the *a* pathway and 0.156 to 0.089 for the *b* pathway. We failed to note any meaningful differences in the bias induced in the *a* and *b* pathways across all sample sizes and effect sizes. The overall bias trend followed a pattern where the bias estimates decreased as the sample size increased. Hence, in the following passage, bias was only reported for the *a* pathways as the differences in bias induced between *a* and *b* pathways were negligible. It is important to note that the absolute bias induced reduced at varying rates across penalties.

The adaptive Lasso penalty was consistently less biased than the Lasso penalty in both *a* and *b* pathways for medium and large effect sizes; however, for the small effect size, the bias induced in the adaptive Lasso was higher when smaller sample sizes were concerned. For M_S, the absolute bias induced for the adaptive Lasso penalty ranged from 0.090 to 0.010. It is important to note that the lower bound of bias (for smaller sample sizes) observed for the adaptive Lasso penalty was greater compared to the Lasso penalty for both pathways. The absolute bias for M_M ranged from 0.065 to 0.029 which was a marked reduction compared to

the Lasso penalty. Similarly, the absolute bias induced for M_L was improved over the Lasso penalty, with estimates ranging from 0.100 to 0.084.

Examining the magnitude of bias for the SCAD and MCP penalties, it was apparent that these two regularization penalties resulted in the lowest scores of absolute bias. However, the absolute bias estimates did not display a linear pattern of reduction as the sample size grew. Instead, we observed similar levels of absolute bias estimates for all sample sizes. For M_S, the absolute bias estimates for the SCAD penalty ranged from 0.014 to 0.003. The absolute bias estimates for the SCAD and MCP penalty were largely equivalent for M_M and M_L. However, for M_S, MCP reported much lower levels of bias compared to SCAD when dealing with sample sizes smaller than 500 with estimates ranging from 0.006 to 0.002. The absolute bias induced in the *b* pathways was consistent with the absolute bias estimates in the *a* pathways for all regularization penalties and sample sizes displaying no overarching patterns.

Chapter 4. Discussion

Methodologists have suggested that there are two primary reasons as to why we should consider implementing regularization methods as an alternative to the least squares estimate in the context of regression (e.g. Hastie et al., 2016; Helwig, 2017). The first reason is prediction accuracy, regularization methods constrain the coefficients of interest towards zero and add some bias to the estimates. Although this may seem counterintuitive, introducing some bias may reduce the variance of the predicted values across samples and improve the overall prediction accuracy and generalizability of the results. The second reason is the ease of interpretation. Certain regularization methods such as the Lasso shrink some estimates exactly to zero, which has the added benefit of simplifying the model. This allows us to identify a sparse model which contains only the coefficients with the largest effects. Jacobucci (2017) extended the use of regularization methods to SEM models with the goal of creating sparse models that are easy to interpret in the context of SEM. Serang and colleagues (2017) later implemented the Lasso penalty to mediation models with the goal of selecting the correct subset of mediators in an exploratory domain. In the current study, Monte Carlo simulation techniques were utilized to compare the variable selection performance of various regularization penalties across a range of sample sizes (50 to 2000). Additionally, the current study compared the relative shrinkage of *a* and *b* pathways and whether that led to changes in variable selection rates.

The overarching goal was to detect the optimal regularization methods that allow us to detect the correct subset of mediators while minimizing Type I error rate. Although each regularization penalty displayed unique patterns of variable selection, each regularization method seemed to approach optimal signal selection asymptotically as the sample size grew to 2000. Simulation conditions for the current study were built upon the framework laid out by Serang and colleagues (2017); however, certain conditions were optimized. Such as the

number of replications, the number of mediators present in the model, effect size used and how the lambda parameter was selected. Since only one regularization penalty (Lasso) was implemented in the original study, we examined whether the current study replicated the results. The Lasso penalty in the current study displayed a similar pattern of variable selection across all samples compared to XMed. However, the Type I error rates for the Lasso increased as the sample size grew larger, which was in direct contrast to the performance of XMed. This was an unexpected result as the Type I error rates reduced for all other penalties in question as the sample size grew. This can be partially explained by the poor consistency of results displayed by the Lasso penalty (Li & Jacobucci, 2021). Moderate Type I error rates were observed for the Lasso penalty across all sample sizes. Although the large Type I error rates were not seen as problematic as the primary goal in an exploratory domain is the primary detection of the correct subset of variables, hence the goal should be to minimize Type II error rates as there is an opportunity to correct Type I errors later on.

The adaptive Lasso was the only method to recover the true model 100% of the time as hypothesized by its oracle property. One of the requirements for the oracle property requires the convergence rate to be optimal. Hence, consistency of model selection is an asymptotic result in adaptive Lasso. The Lasso penalty can be seen as liberal and often includes more variables than the optimal model. Consequently, the true model is very likely a subset of these variables. Therefore, using a secondary estimation stage like the adaptive Lasso helps us achieve more consistent selection rates. For smaller sample sizes, the SCAD penalty performed similarly to the Lasso penalty, reporting moderate to large Type I error rates, and mimicked the performance of adaptive Lasso for larger samples with low Type I error rates. Hence, we could consider the penalties on a continuum, with MCP generally being the most liberal penalty, selecting both the signal mediators and the noise variables in most replications. In comparison, the adaptive Lasso can be considered more conservative as

it is much more likely to filter out mediators. The Lasso penalty is in the middle of the continuum with moderate selection rates and moderate Type I error rates. However, it is important to consider that the selection rates for each regularization method can be further optimized if alternative methods are chosen for selecting the tuning parameter.

As expected, no single regularization method consistently outperformed the others across all sample sizes, and each method posed a unique solution that may be tailored to a researcher's interests. For example, the MCP penalty would be an appropriate choice if the goal is to correctly identify the correct subset of mediators in an exploratory domain with no regard for Type I error rates. Similarly, if the primary goal is to correctly select the correct subset of mediators whilst minimizing Type I error rates and the research in question falls in the middle of the exploratory-confirmatory continuum a Lasso, SCAD or adaptive Lasso penalty would be more appropriate choices. However, if the sample size for the hypothesized study in question approaches 1000, adaptive Lasso and the SCAD penalty identify the true model characteristics with near certainty. The results from the current simulation study largely align with another study that demonstrated the inconsistency and high Type I error rates displayed by the Lasso penalty (Li & Jacobucci, 2021; Serang et al., 2017). The authors further recommended the use of stability selection when applying regularization methods in the context of SEM. Stability selection for regularization methods allows us to improve variable selection performance by aggregating the selection results from bootstrap samples. Furthermore, It provides sample control for false positive rates and enables us to choose an appropriate value of lambda (Li & Jacobucci, 2021)

Another goal of the study was to compare the selection rates of M_{SL} and M_{LS}. It was initially hypothesized that both the *a* and *b* pathways would be shrunk similarly towards zero, and no patterns would emerge. We discovered that the two mediators were being selected at different rates for the Lasso and SCAD penalty, indicating that they were being shrunk

disproportionately. This has major implications when the product of coefficients approach is employed in testing mediation models, as one pathway being shrunk to zero before the other filters out the mediator from the model. Since the current study employed the product of coefficients approach to estimating indirect effects, this highlights a cause for concern. In recent literature, Kesteren and Oberski (2019) demonstrated that the regularization penalty can be built into the mediation model to penalize the indirect effects as a whole instead of individual *a* and *b* pathways. They named this novel method the Coordinate-Wise Mediation Filter (CMF). This procedure used a cyclical coordinate descent algorithm and applied a decision function to the mediators conditional upon the set of currently selected mediators. The decision function represented a dichotomous decision (whether the proposed mediator in question is selected or not) based on the estimate of the indirect effect.

The final part of the study examined the absolute bias induced in the *a* and *b* pathways for M_S, M_M and M_L. Although the disproportionate shrinkage indicated that the *a* and *b* pathways were penalized disproportionately. It is hard to justify that line of reasoning when the absolute bias induced in the *a* and *b* pathways was similar across all sample sizes for each regularization penalty. Adaptive Lasso, MCP and SCAD penalties were all superior to the Lasso at minimizing the induced bias. However, if the goal is to minimize the induced bias, relaxed Lasso has been shown to be the most appropriate option (Serang et al., 2017).

4.1 Implications

The seminal part of any research design involves the construction of theory. Haig (2014) described that theory construction occurs in three stages. The first stage involves theory generation and is most commonly associated with exploratory research. The third and final stage involves theory appraisal and is most commonly implemented in SEM and psychological research. This stage represents the confirmatory end of the spectrum. The second stage relates to theory development and is often represented in the middle of the

exploratory-confirmatory domain. Brandmaier and Jacobucci (2022) described this stage as a middle grey area in the construction of theory. The authors further suggested that adopting specific ideas from machine learning research would permit us to forgo the idea that we can specify a true model and instead allow us to blend confirmatory approaches with more exploratory research. This would aid us in fitting models that predict well and are easier to interpret. This way, we may approach a medium between a purely theory-driven model and the best fitting model, which may be in the middle of the exploratory-continuum domain instead of an extreme. The utility of RegSEM in the context of mediation models is maximized when available theory is limited or when available theory fails to identify certain aspects that were erroneously included in the model.

The stage of theory development involves assessing confounding variables along with other features of the data that provide us with possible solutions that allow us to expand or amend existing theory. The primary concern when fitting an exploratory model is that of overfitting (or equivalently, a lack of generalizability) or exhausting available degrees of freedom if the number of predictors is high (Brandmaier & Jacobucci, 2022). However, adopting techniques from machine learning literature has allowed us to identify algorithms that tend to overfit the sample data and has provided us with countermeasures to both assess and prevent overfitting. One lesser known advantage of regularization methods compared to other variable selection methodologies relates to the use of degrees of freedom. Other methods tend to use more than *P* (predictors) degrees of freedom; however, regularization methods allow us to count the degrees of freedom by the number of non-zero coefficients in the model (Hastie et al., 2016). In the context of linear models, it has been demonstrated that the degrees of freedom are equal to the number of predictors in the model. However, under adaptive fitting procedures, it is typically the case that the degrees of freedom are larger than *P*. Research has shown that the Lasso penalty uses the degrees of freedom that are equal to

the number of non-zero coefficients in the model (Zou, Hastie & Tibshirani, 2007; Tibshirani & Taylor, 2012).

4.2 Limitations and Directions for Future Research

This study has several limitations which can serve as future research directions. Before addressing such limitations, it is essential to note that the current study dealt with a linear causal model. Although mediation and causation do not need to be linear, the use of SEM imposes a linearity assumption on the mediators. (Chen, Gunzler, Wu & Zhang, 2013). The current study penalized individual regression pathways instead of indirect effects as a whole which may have led to disproportionate selection rates. Future studies should compare the performance of CMF to RegSEM with a variety of simulation factors to examine if this trend persists. Furthermore, the current study highlighted the disproportionate selection rates of the *a* and *b* pathways using alternative mediators, a better alternative would be to count the proportion of replications where the *a* pathway and the *b* pathway were shrunk to zero.

The optimal selection rates of the Lasso penalty and its generalizations largely depend on the selected value of the penalty term. Hence, it is of the utmost importance that future studies examine the optimal way of choosing a penalty term that provides us with a sparse solution without overly penalizing the estimates. Additionally, there are several simulation factors that could be included in the simulation design to enhance our understanding of regularization methods in mediation models. For one, we could use the total number of mediators to examine high-dimensional setups. The Lasso penalty has been demonstrated to be particularly effective in these settings. Second, we could free the residual covariances between mediators to explore how that would impact selection rates. And finally, we could impose residual correlational structures among the variables. Other regularization penalties such as elastic net and Ridge regression have been demonstrated to work well when

predictors are correlated to each other (Zou & Hastie, 2005). This warrants further investigation and could serve as an inspiration for future studies.

Li and Jacobucci (2021) highlighted the advantages of using stability selection to determine the range of lambda values that may allow us to determine the optimal penalty. The current study relied on standards established by Jacobucci (2017) and used minimum values of BIC to determine the optimal lambda penalty. Researchers should compare the performance of different fit indices in determining the optimal lambda penalty or use stability selection to further refine their choice. Similarly, we relied on the upsilon measure to determine the benchmarks for small, medium and large effect sizes; however, it is important to note there is a lack of literature that determines the optimal effect sizes in the case of multiple mediation models (Lachowicz, Preacher & Kelley, 2018).

Finally, it would behoove researchers to extend regularization methods to more complex models. For example, we could include multiple predictors and outcome variables along with moderators. The current study primarily addressed additive causal models; the additivity assumption means that the association between the predictor and the outcome does not depend on other predictors or mediators in the model. Future studies should relax this additivity assumption. The addition of such components would not only allow us to address more complex models but also decrease the available degrees of freedom, enhancing the need for variable selection. This would make the inclusion of regularization procedures all the more suitable. Additionally, we could examine the performance in the context of serial mediation models. It is important to note that RegSEM allows us to include latent variables and determine which parameters to penalize so researchers can penalize specific parameters in the model that are not determined by theory. Future researchers should also seek to adapt regularization methods to growth curve models or multilevel models in the context of SEM.

Chapter 5. Conclusions

A series of Monte Carlo simulations indicated that there is no free lunch when applying different regularization penalties in the context of mediation models. No single method consistently outperformed the others for all simulation conditions. The current study allows researchers to make an informed decision about the regularization penalty they would like to implement while seeking the relevant set of mediators that should be included in the final model. Furthermore, researchers are always encouraged to only penalize variables where existing theory is limited as it is not necessary to penalize all selected mediators. The study attempts to draw researchers' attention to the possible benefits and concerns of applying various regularization penalties in the context of a mediation model. Furthermore, the simulation study highlights the disproportionate selection rates for certain mediators although the *a* and *b* pathways were being shrunk similarly. This may be in part due to the problems associated with penalizing individual pathways instead of indirect effects as a whole. Researchers who wish to implement regularization procedures for the purposes of variable selection should additionally ensure that they utilize stability selection procedures to optimize their choice of penalty and not base the decision simply on fit indices.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 37-47.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine,* 66, 411–421

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology,* 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173

Brandmaier, A. M., & Jacobucci, R. C. (2022). Machine-learning approaches to structural equation modeling. *In Handbook of structural equation modeling*. Guilford Press.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373-384.

Cheung, M. W. (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods*, 41, 425–438. http://dx.doi.org/10.3758/BRM.41.2.425

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Fairchild, A. J., MacKinnon, D. P., Taborga, M. P., & Taylor, A. B. (2009). R^2 effect-size measures for mediation analysis. *Behavior Research Methods*, 41, 486–498. http://dx.doi.org/10.3758/BRM.41.2.486

Fairchild, A. J., & McDaniel, H. L. (2017). Best (but oft-forgotten) practices: mediation analysis. *The American Journal of Clinical nutrition*, *105*(6), 1259–1271. https://doi.org/10.3945/ajcn.117.152546

Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group Lasso and a sparse group Lasso. *arXiv preprint arXiv:1001.0736*.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, Massachusetts: MIT press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning.* New York: SpringerVerlag

Hastie, T., Tibshirani, R., & Wainwright, M. (2016). *Statistical learning with sparsity: the Lasso and generalizations*. Chapman and Hall/CRC.

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology*, *13*(1), 1-19.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys*, *2*, 61-93.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67. doi:10.2307/1267351

Huang, J., Ma, S., & Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 1603-1618.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, 23(4), 555-566.

Jacobucci, R. (2017). Regsem: Regularized structural equation modeling. *arXiv Preprint arXiv:1703.08489.*

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602-619.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17, 137–152. http://dx.doi.org/10.1037/a0028086

Gunzler, D., Chen, T., Wu, P., & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry*, 25(6), 390–394. https://doi.org/10.3969/j.issn.1002-0829.2013.06.009

Lachowicz, M. J., Preacher, K. J., & Kelley, K. (2018). A novel measure of effect size for mediation analysis. *Psychological Methods*, *23*(2), 244.

Li, X., & Jacobucci, R. (2021). Regularized structural equation modeling with stability selection. *Psychological Methods.*

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the Lasso. *The Annals of Statistics*, 42, 413–468. doi:10.1214/13-aos1175

MacKinnon, D. P. (2012). *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*(2), 144-158.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, *7*(1), 83-94.

McNeish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, 50, 471–484. doi:10.1080/00273171.2015.1036965

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, *52*(1), 374-393.

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research,* 44(3), 692-716.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879-891.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93–115. doi:10.1037/a0022658

R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36.

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory Mediation analysis via regularization. *Structural Equation Modeling*, 24, 733–744. doi:10.1080/10705511.2017.1311775

Serang, S., & Jacobucci, R. (2020). Exploratory mediation analysis of dichotomous outcomes via regularization. *Multivariate Behavioral Research*, *55*(1), 69-86.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *7*(4), 422.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290-312.

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

Tibshirani, R. and Taylor, J. (2012), Degrees of freedom in Lasso problems, *Annals of Statistics* 40(2), 1198–1232.

Van Kesteren, E. J., & Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 710-723.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* (Cambridge, Mass.), 21(4), 540.

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3), 553–568.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49-67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. T*he Annals of Statistics*, 38(2), 894–942.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society,* Series B, 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007), On the degrees of freedom of the Lasso, *Annals of Statistics* 35(5), 2173–2192.

Table 1

Frequency of converged and proper solutions for each design row

Sample Size	Converged	Percent Converged
	LASSO	
50	994	99.4
100	1000	100.0
200	1000	100.0
500	1000	100.0
1000	1000	100.0
2000	1000	100.0
	Adaptive LASSO	
50	989	98.9
100	1000	100.0
200	1000	100.0
500	1000	100.0
1000	1000	100.0
2000	1000	100.0
	SCAD	
50	996	99.6
100	1000	100.0
200	1000	100.0
500	1000	100.0
1000	1000	100.0
2000	1000	100.0

	MCP	
50	990	99
100	1000	100
200	1000	100
500	1000	100
1000	1000	100
2000	1000	100

Table 2
Absolute bias induced in the Lasso and adaptive Lasso penalty

	Small		Medium		Large	
Sample Size	**bias x**	**bias y**	**bias x**	**bias y**	**bias x**	**bias y**
			LASSO			
100	0.067	0.062	0.105	0.101	0.152	0.156
200	0.057	0.062	0.091	0.094	0.132	0.136
300	0.038	0.034	0.067	0.060	0.113	0.117
500	0.033	0.034	0.053	0.055	0.106	0.101
1000	0.025	0.032	0.047	0.048	0.099	0.103
2000	0.018	0.017	0.042	0.037	0.095	0.089
			ADAPTIVE LASSO			
100	0.090	0.086	0.065	0.069	0.100	0.097
200	0.071	0.065	0.045	0.044	0.096	0.100
300	0.057	0.064	0.039	0.034	0.085	0.089
500	0.031	0.032	0.037	0.044	0.086	0.084
1000	0.016	0.013	0.029	0.024	0.084	0.092
2000	0.010	0.009	0.029	0.030	0.084	0.076

Table 3
Absolute bias induced in the SCAD and MCP penalty

	Small		Medium		Large	
Sample Size	bias x	bias y	bias x	bias y	bias x	bias y
			SCAD			
100	0.014	0.013	0.029	0.032	0.082	0.088
200	0.018	0.017	0.031	0.037	0.085	0.088
300	0.015	0.013	0.028	0.026	0.082	0.075
500	0.012	0.016	0.028	0.024	0.082	0.076
1000	0.004	0.002	0.026	0.034	0.083	0.091
2000	0.003	0.005	0.027	0.029	0.082	0.083
			MCP			
100	0.006	0.005	0.030	0.037	0.083	0.075
200	0.002	0.007	0.026	0.023	0.083	0.085
300	0.001	0.000	0.025	0.029	0.085	0.093
500	0.001	0.008	0.026	0.026	0.081	0.075
1000	0.002	0.002	0.028	0.027	0.082	0.074
2000	0.002	0.009	0.026	0.033	0.082	0.076

Table 4
Variable Selection rates for the Lasso and adaptive Lasso penalty

	Signal			Noise		X and Y paths	
Sample Size	Small	Medium	Large	Noise	Noise	Small X | Large Y	Large X | Small Y
				LASSO			
50	0.299	0.871	0.990	0.194	0.174	0.457	0.590
100	0.450	0.979	1.000	0.203	0.175	0.613	0.699
200	0.695	1.000	1.000	0.212	0.193	0.776	0.841
500	0.951	1.000	1.000	0.244	0.274	0.963	0.986
1000	0.999	1.000	1.000	0.237	0.255	1.000	1.000
2000	1.000	1.000	1.000	0.269	0.267	1.000	1.000
				ADAPTIVE LASSO			
50	0.098	0.739	0.979	0.026	0.022	0.271	0.288
100	0.127	0.937	1.000	0.009	0.006	0.321	0.379
200	0.229	0.999	1.000	0.002	0.002	0.478	0.462
500	0.701	1.000	1.000	0.002	0.001	0.783	0.822
1000	0.944	1.000	1.000	0.001	0.005	0.968	0.976
2000	1.000	1.000	1.000	0.000	0.000	1.000	0.998

Table 5
Variable Selection rates for the Lasso and adaptive Lasso penalty

	Signal			Noise		X and Y paths	
Sample Size	Small	Medium	Large	Noise	Noise	Small X | Large Y	Large X | Small Y
				SCAD			
50	0.388	0.927	0.997	0.204	0.222	0.626	0.587
100	0.371	0.995	1.000	0.067	0.062	0.641	0.497
200	0.375	1.000	1.000	0.012	0.012	0.717	0.556
500	0.600	1.000	1.000	0.000	0.002	0.852	0.744
1000	0.928	1.000	1.000	0.002	0.003	0.974	0.966
2000	0.999	1.000	1.000	0.000	0.003	1.000	1.000
				MCP			
50	0.884	0.997	1.000	0.835	0.846	0.954	0.951
100	0.902	1.000	1.000	0.788	0.736	0.939	0.940
200	0.942	1.000	1.000	0.662	0.688	0.980	0.959
500	0.992	1.000	1.000	0.481	0.502	0.997	0.995
1000	1.000	1.000	1.000	0.292	0.305	1.000	1.000
2000	1.000	1.000	1.000	0.156	0.137	1.000	1.000

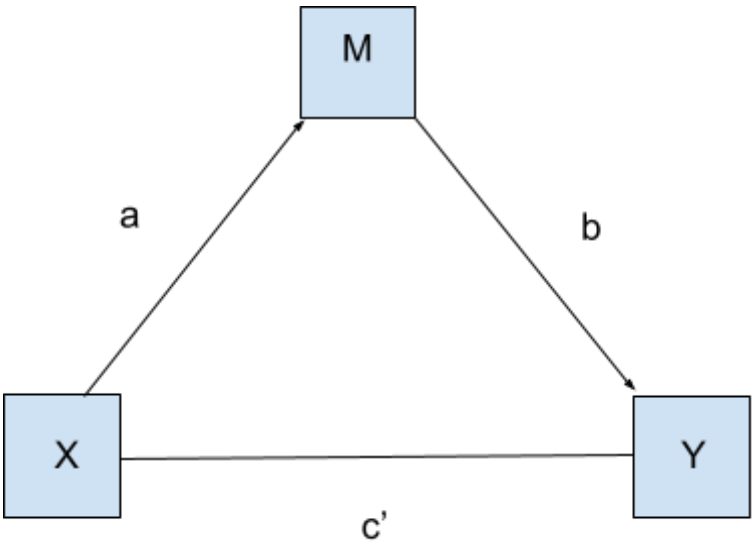

Figure 1. An exemplary model with a single mediator

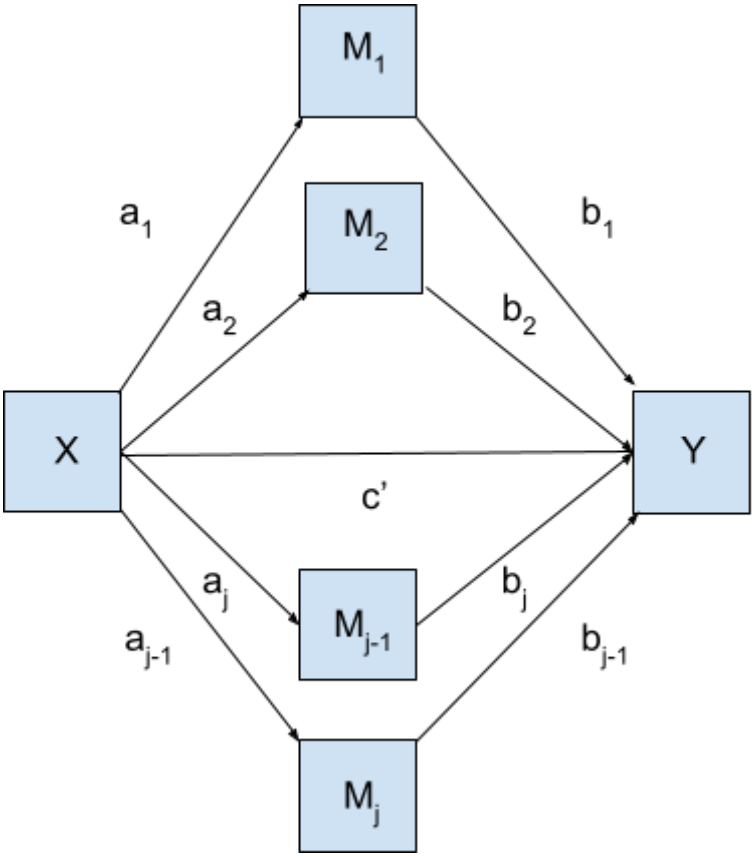

Figure 2. Multiple Mediation Model

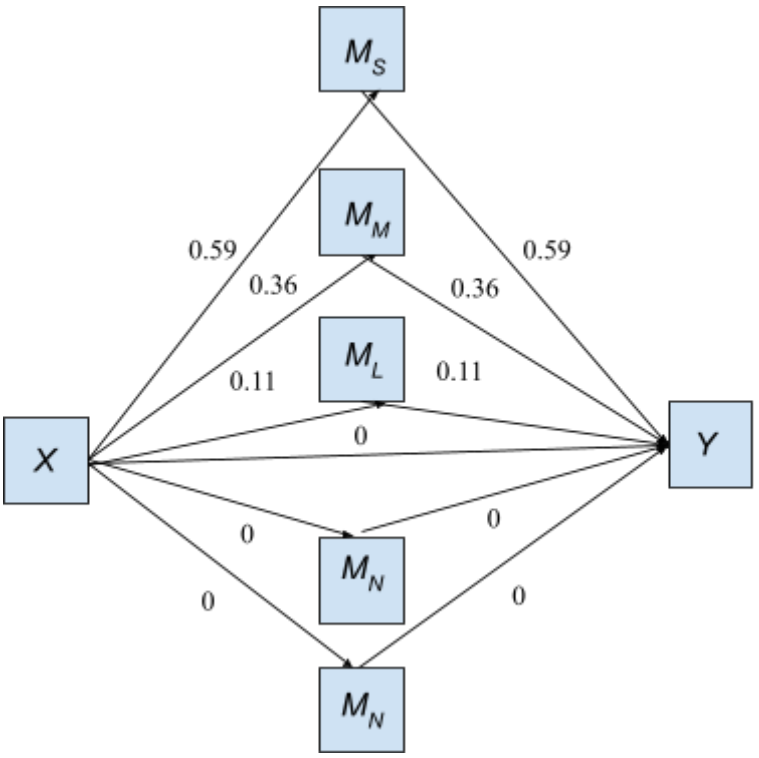

Figure 3. XMed Simulation Design (Serang et al., 2017) with three signal mediators represented by M_S, M_M and M_L and two noise mediators represented by M_N.

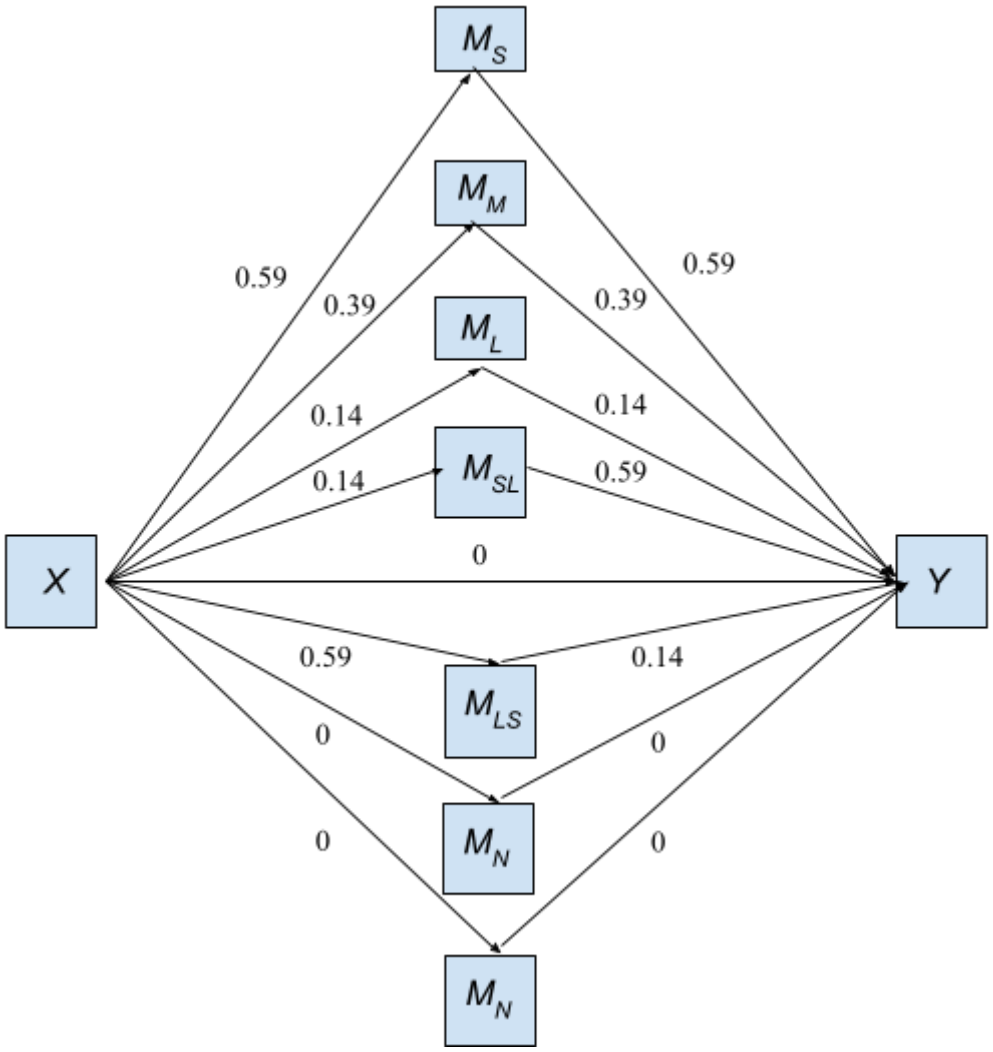

Figure 4. Simulation Design with five signal mediators represented by M_S, M_M, M_L, M_{SL} and M_{LS} and two noise mediators M_N. The *a* pathways are represented by the *X* to *M* regression coefficients whereas the *b* pathways are represented by *M* to *Y* regression coefficients.

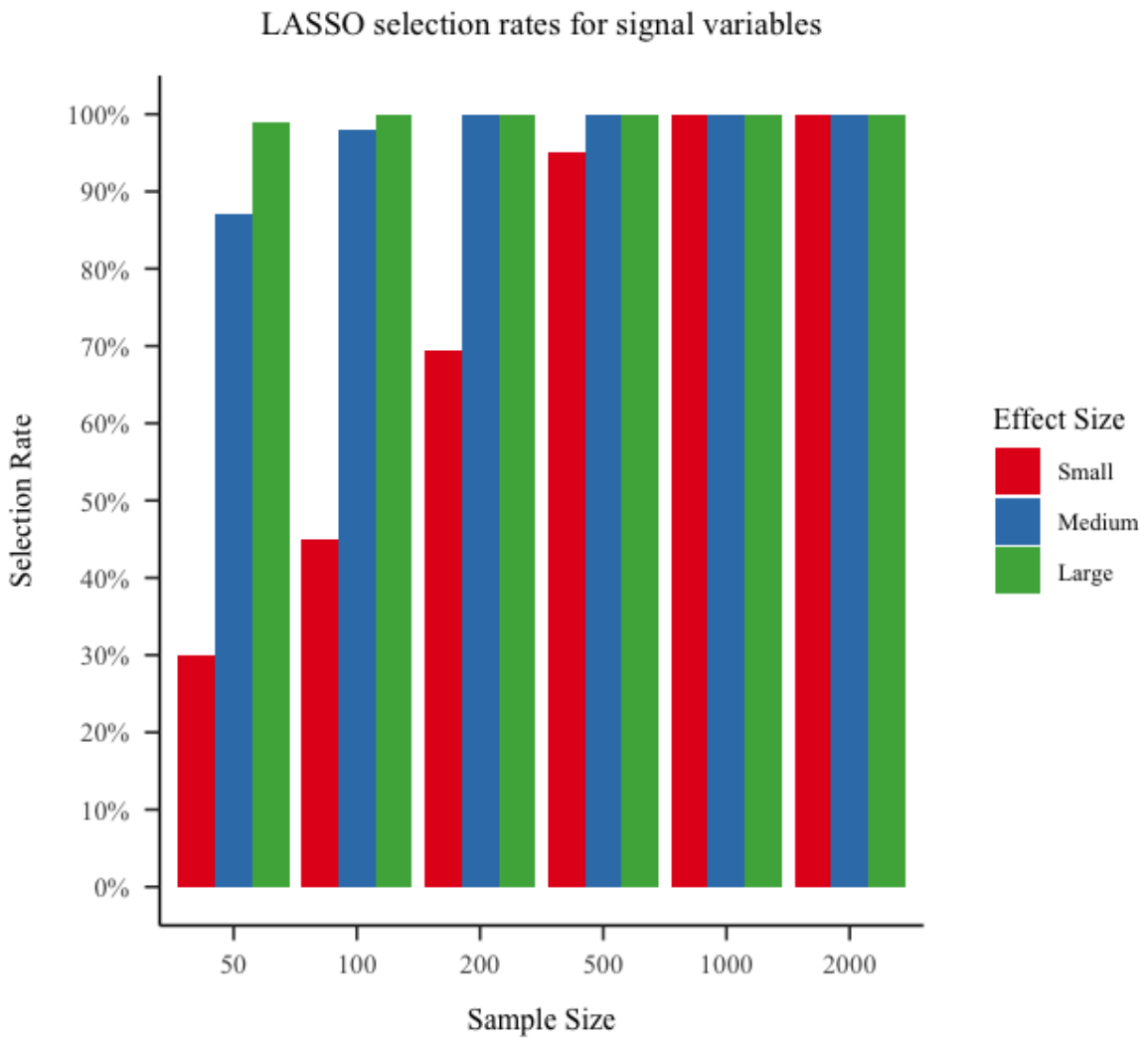

Figure 5. Barplots comparing the selection rates of the three signal variables (M_S, M_M, M_L) for the Lasso penalty across a range of sample sizes.

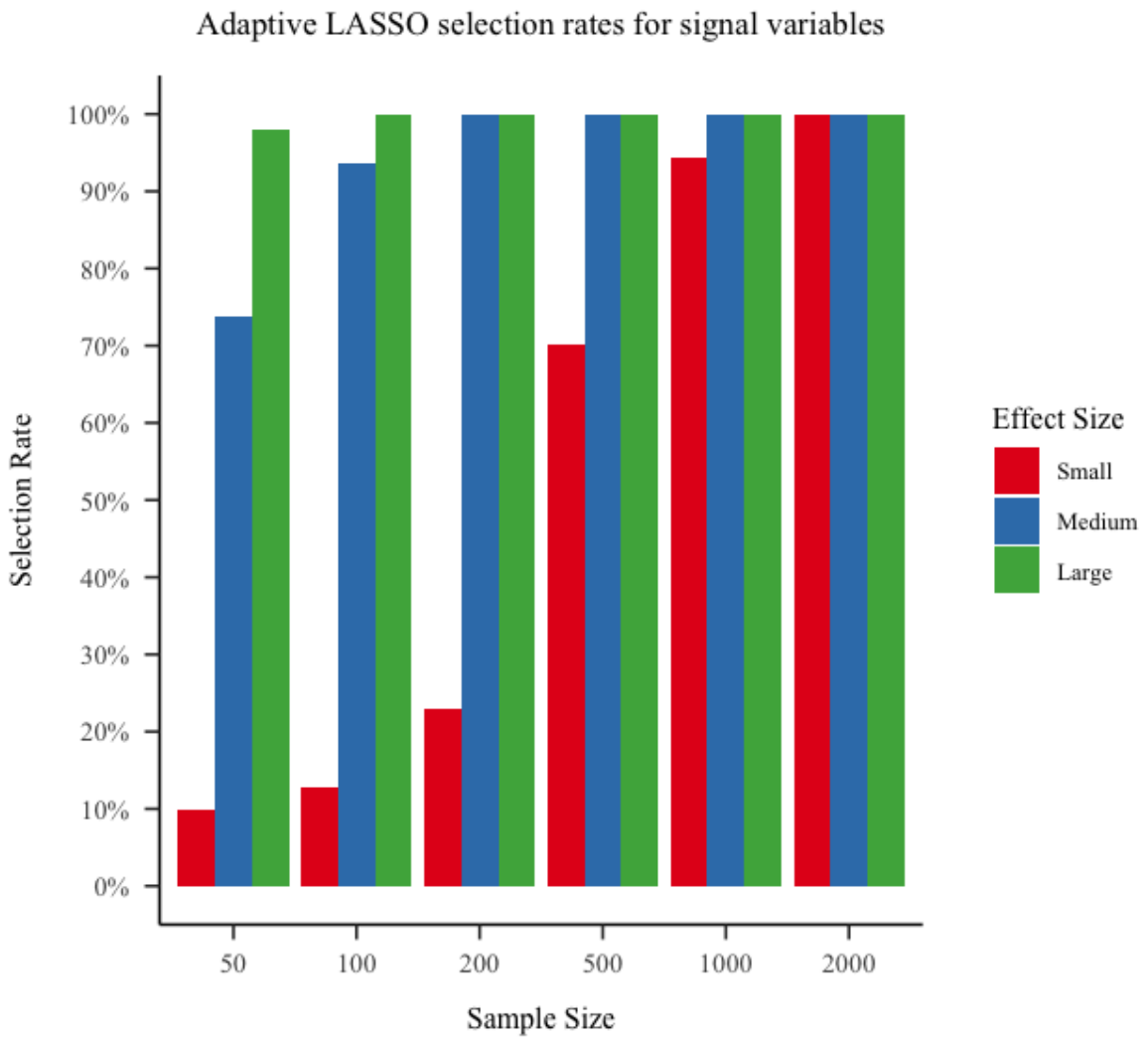

Figure 6. Barplots comparing the selection rates of the three signal variables (M_S, M_M, M_L) for the adaptive Lasso penalty across a range of sample sizes.

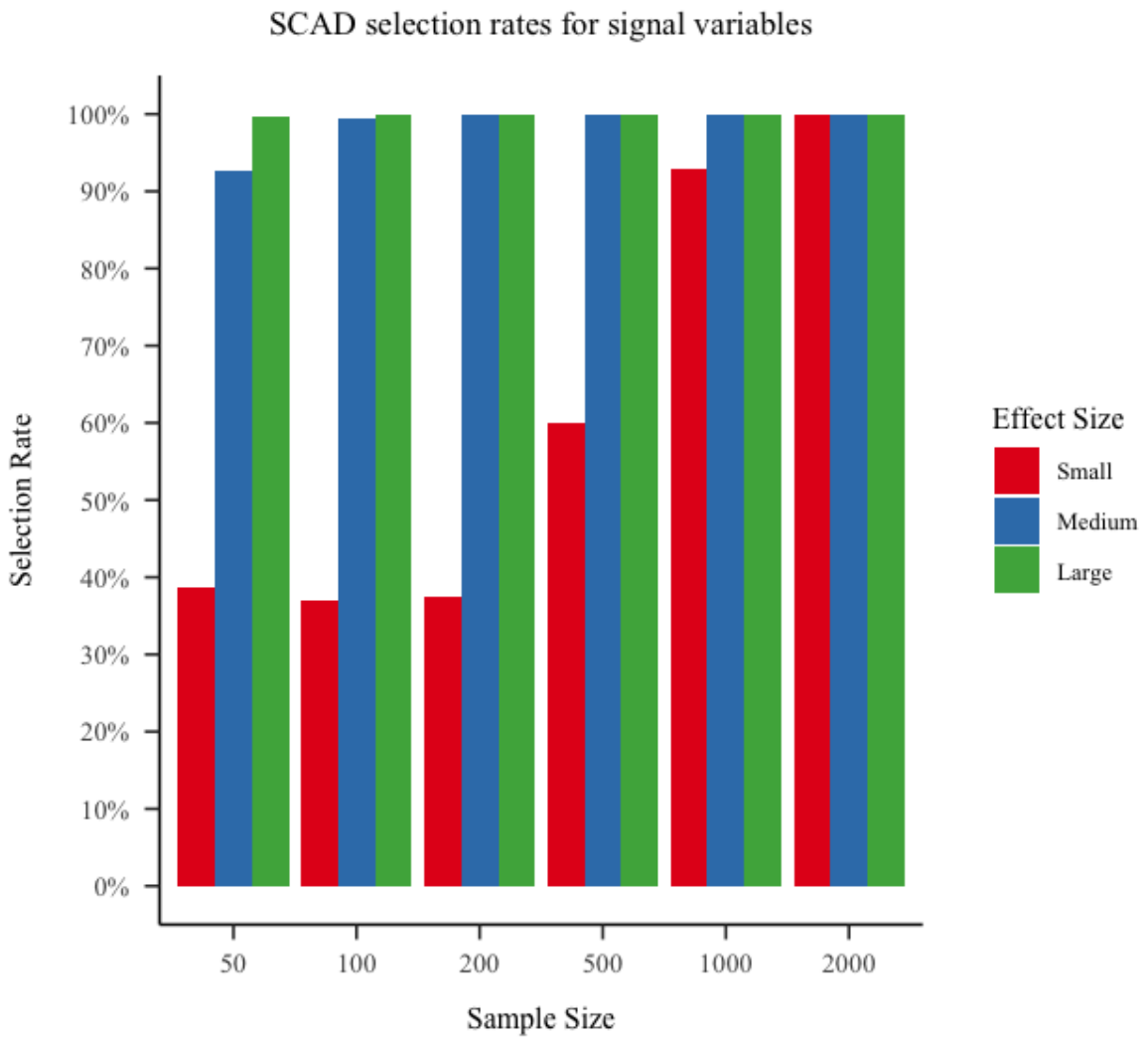

Figure 7. Barplots comparing the selection rates of the three signal variables (M_S, M_M, M_L) for the SCAD penalty across a range of sample sizes.

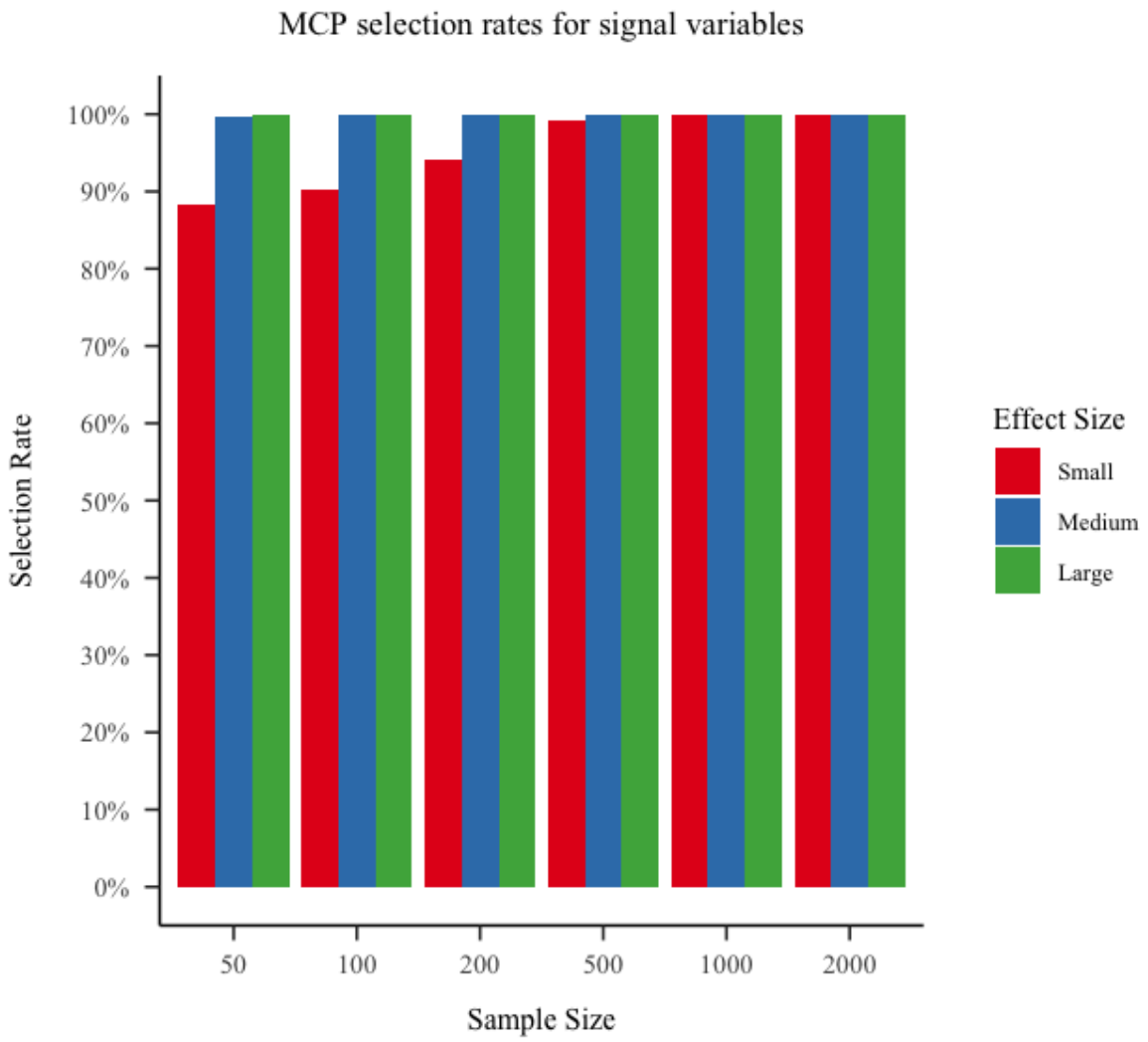

Figure 8. Barplots comparing the selection rates of the three signal variables (M_S, M_M, M_L) for the MCP penalty across a range of sample sizes.

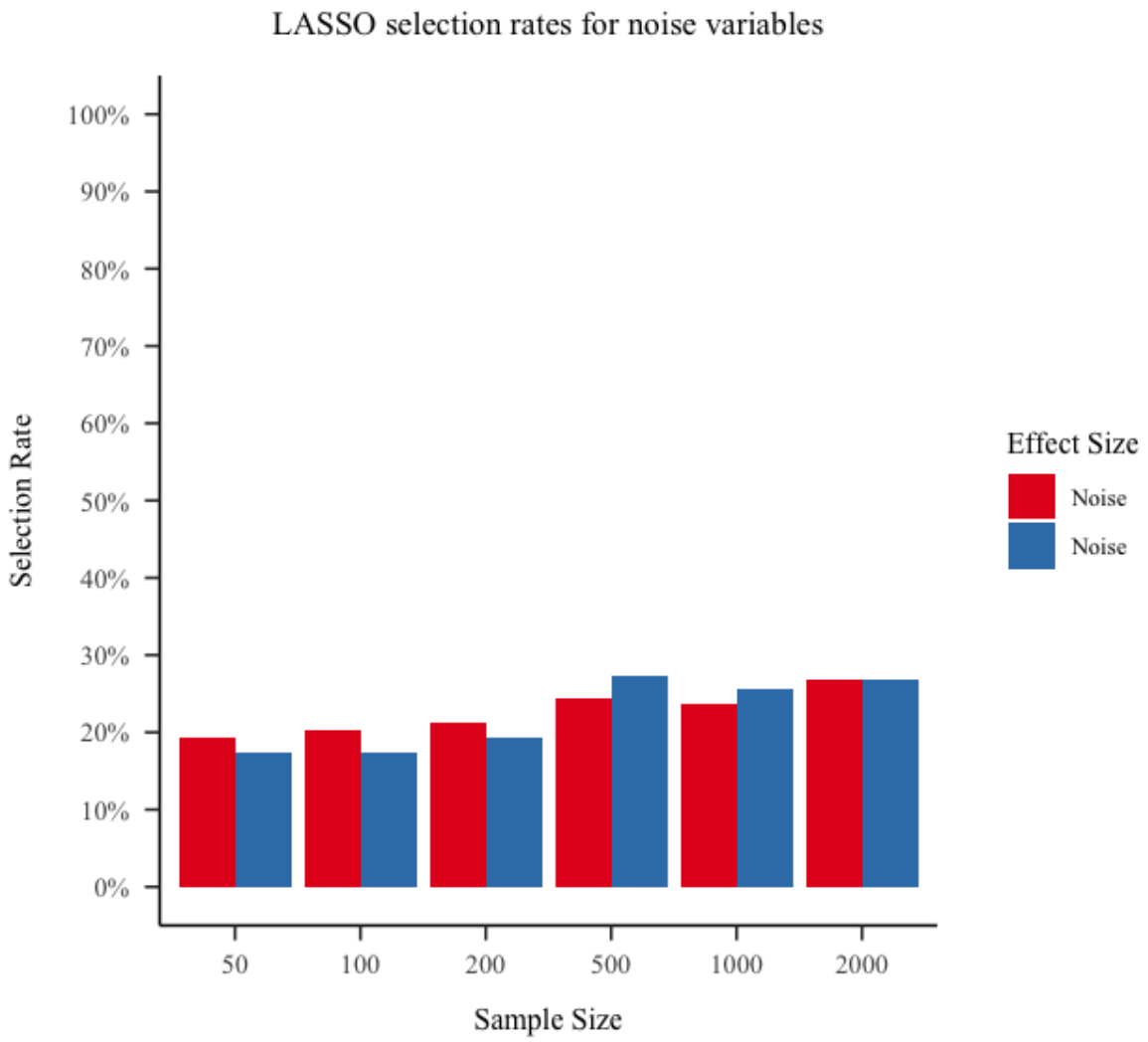

Figure 9. Barplots comparing the Type I error rates of the two noise variables (M_N, M_N) for the Lasso penalty across a range of sample sizes. The Lasso penalty has fairly high Type I error rates that increase as sample size grows larger.

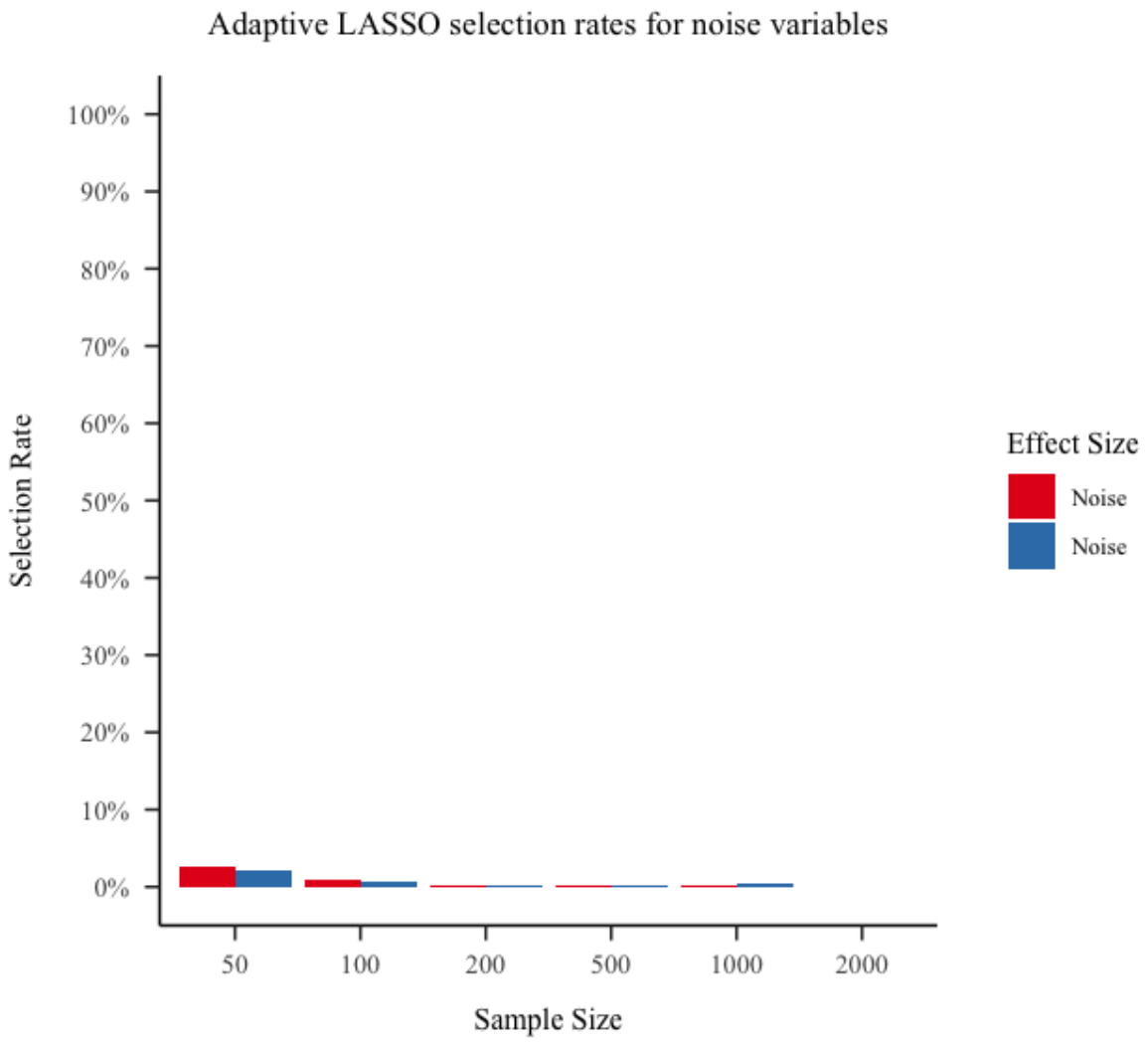

Figure 10. Barplots comparing the Type I error rates of the two noise variables (M_N, M_N) for the adaptive Lasso penalty across a range of sample sizes. The adaptive Lasso penalty has negligible Type I error rates. The scale of the *Y* axis was not changed to highlight the differences across penalties.

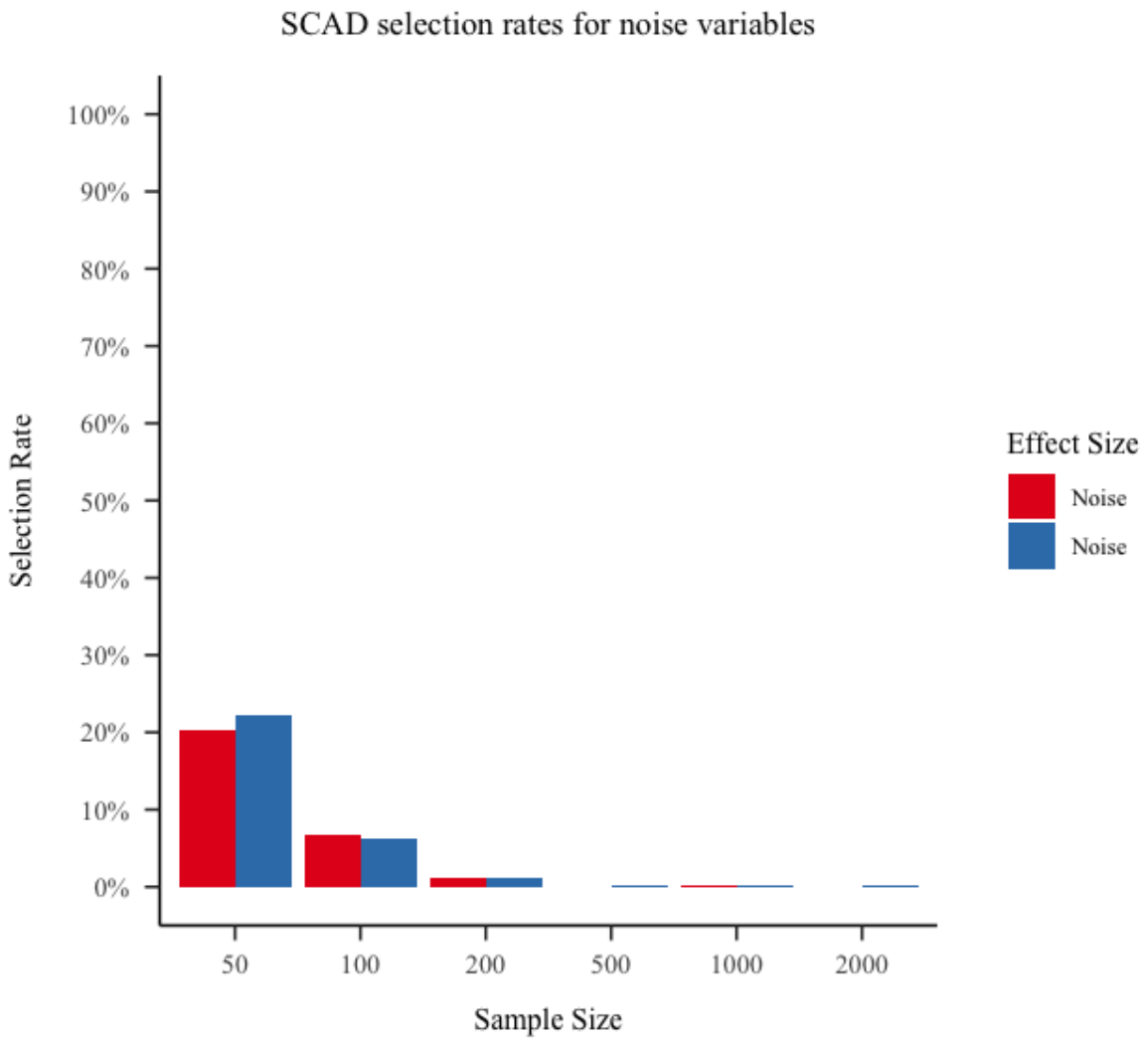

Figure 11. Barplots comparing the Type I error rates of the two noise variables (M_N, M_N) for the SCAD penalty across a range of sample sizes. The SCAD penalty has fairly Type I error rates for smaller samples. But as sample size grows larger the Type I error rates are negligible.

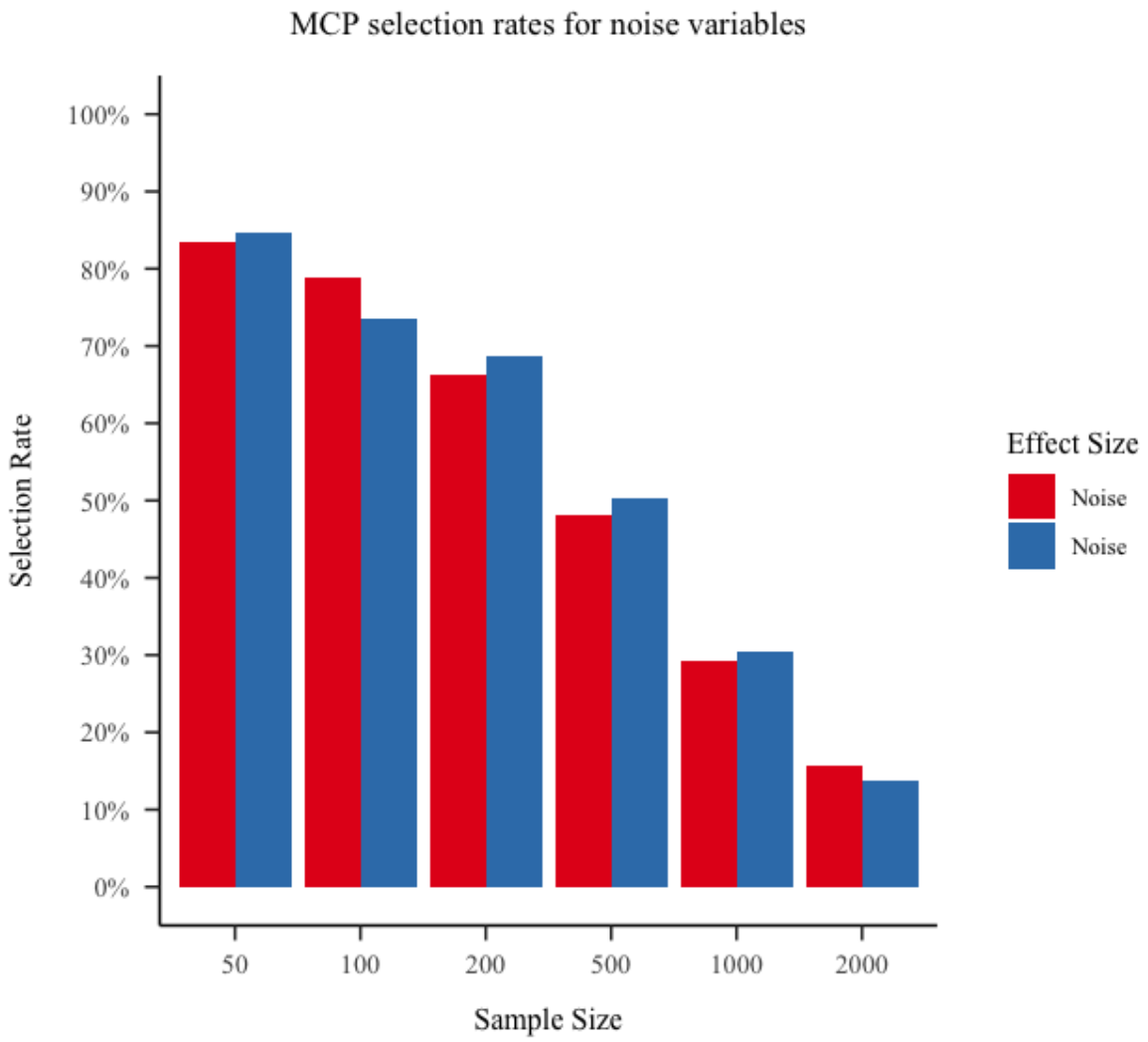

Figure 12. Barplots comparing the Type I error rates of the two noise variables (M_N, M_N) for the MCP penalty across a range of sample sizes. The MCP penalty has extremely high Type I error rates for smaller samples. But as sample size grows larger the Type I error rates decrease substantially.

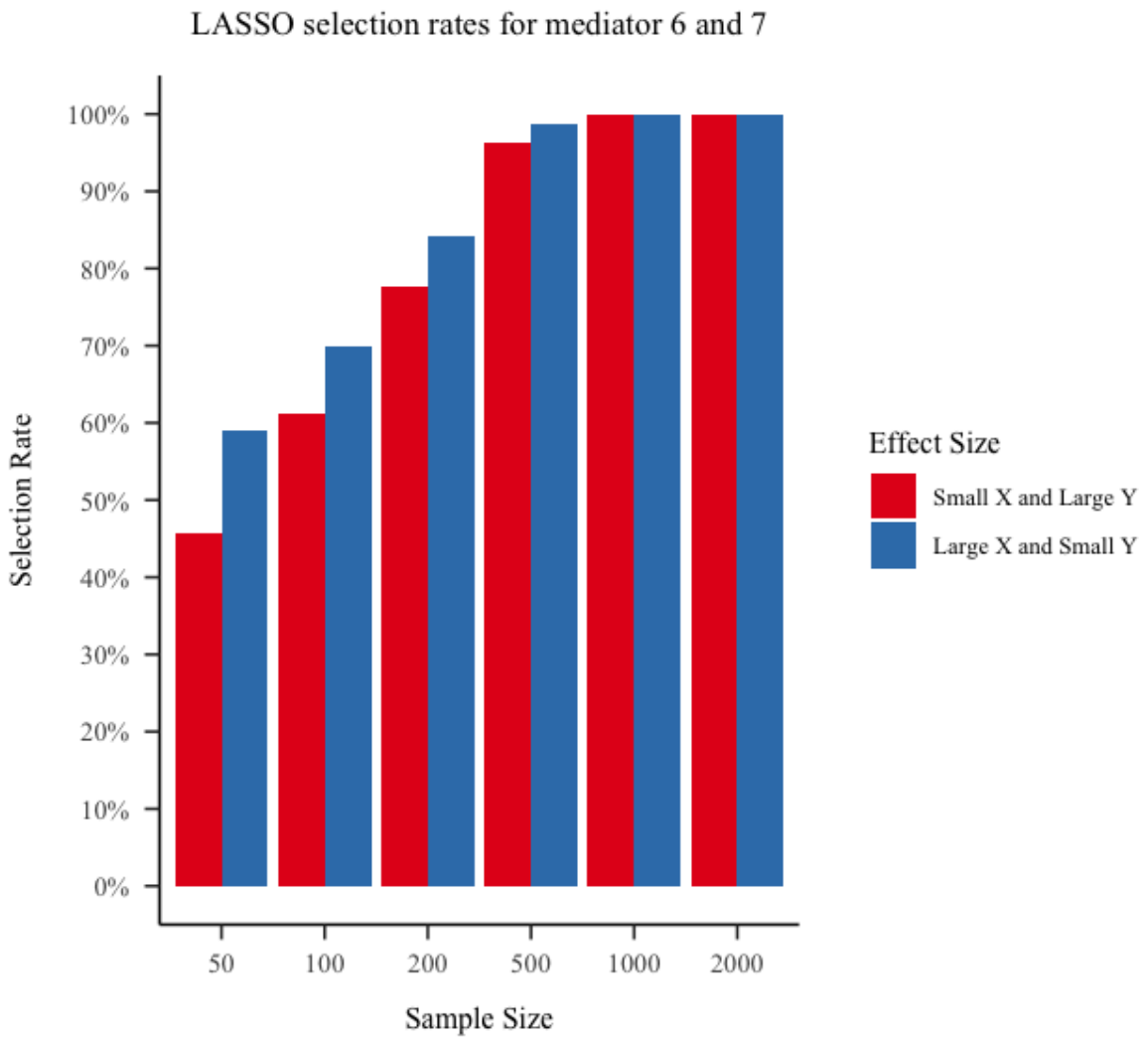

Figure 13. Barplots comparing the selection rates of the two mediators with opposing *a* and *b* pathways (M_{SL}, M_{LS}) for the Lasso penalty across a range of sample sizes. The M_{LS} mediator had higher selection rates for sample sizes upto 500 after eventually levelling off.

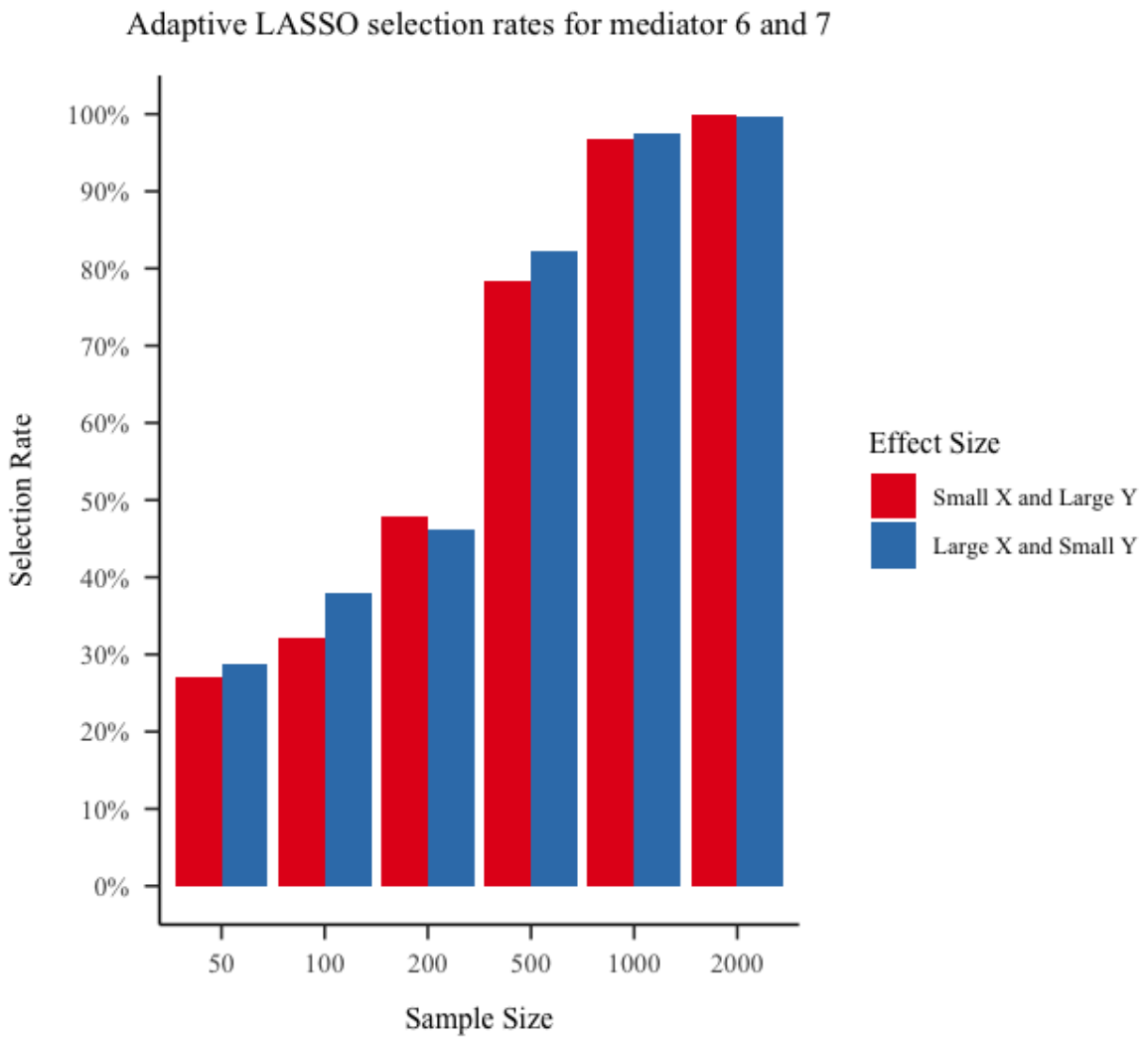

Figure 14. Barplots comparing the selection rates of the two mediators with opposing *a* and *b* pathways (M_{SL}, M_{LS}) for the adaptive Lasso penalty across a range of sample sizes. There was no discernible pattern of differences in selection rates for the two mediators.

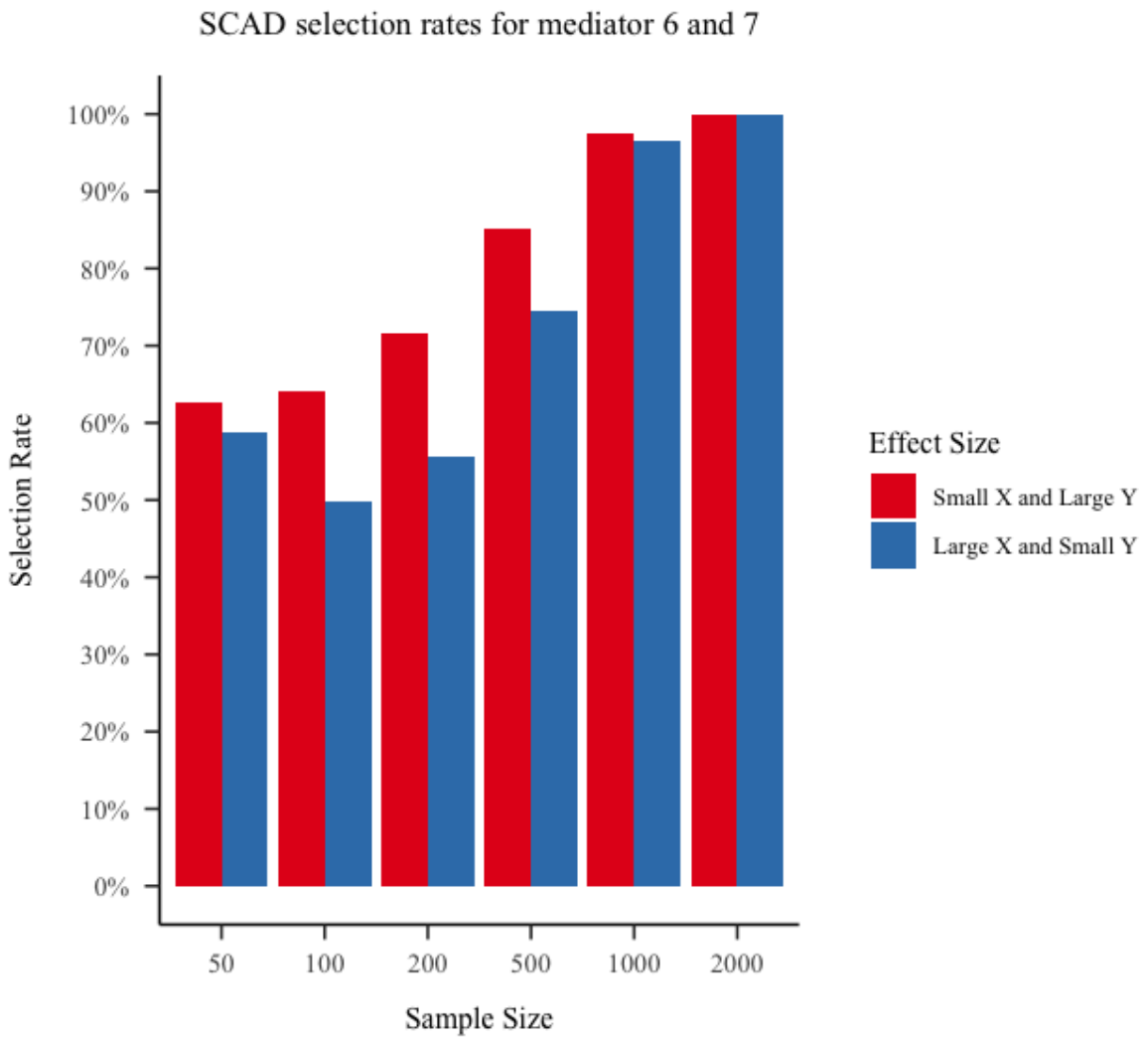

Figure 15. Barplots comparing the selection rates of the two mediators with opposing *a* and *b* pathways (M_{SL}, M_{LS}) for the SCAD penalty across a range of sample sizes. The M_{SL} mediator had considerably higher selection rates for sample sizes upto 500 after eventually levelling off.

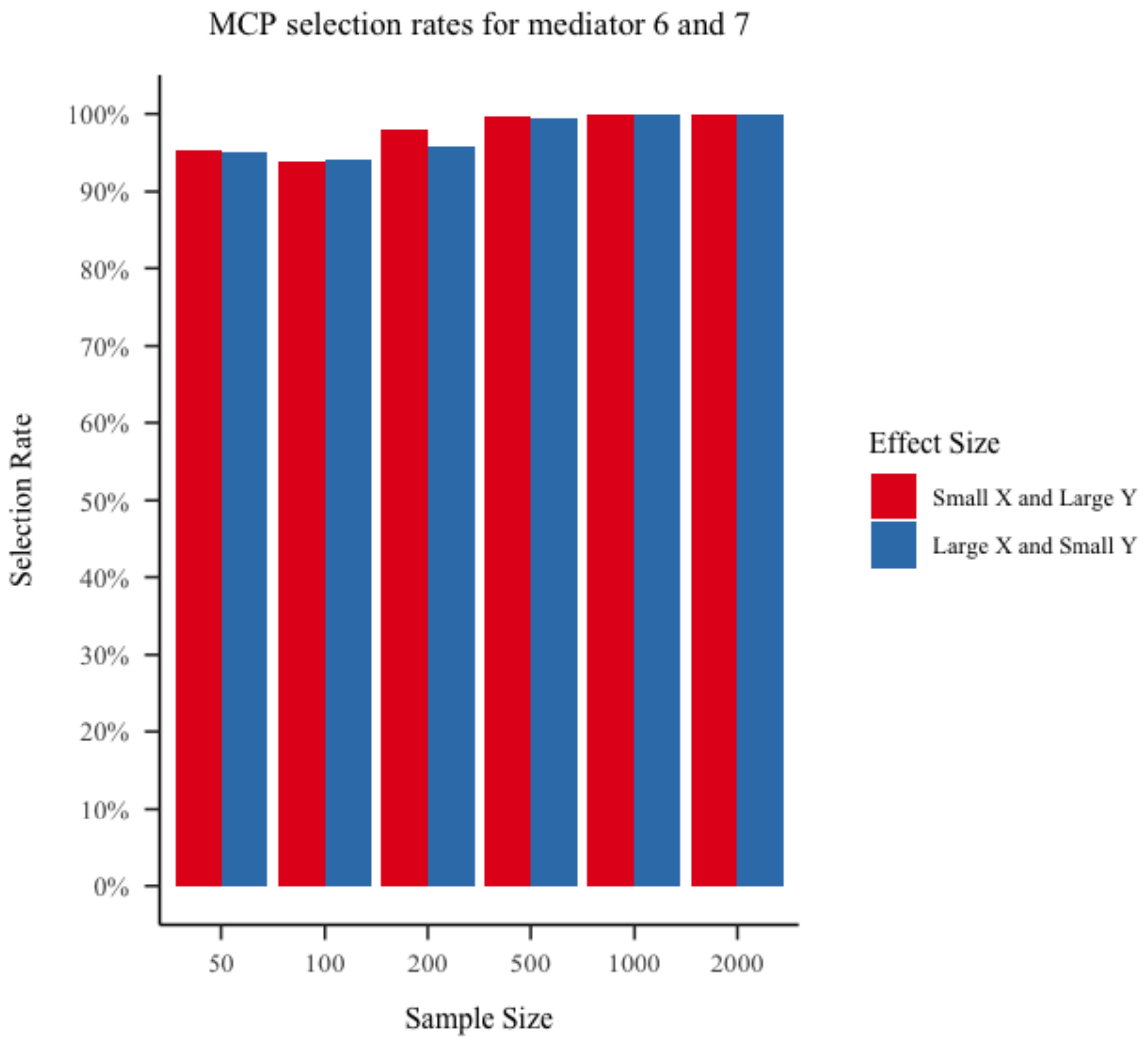

Figure 16. Barplots comparing the selection rates of the two mediators with opposing *a* and *b* pathways (M_{SL}, M_{LS}) for the MCP penalty across a range of sample sizes. There was no discernible pattern of differences in selection rates for the two mediators.

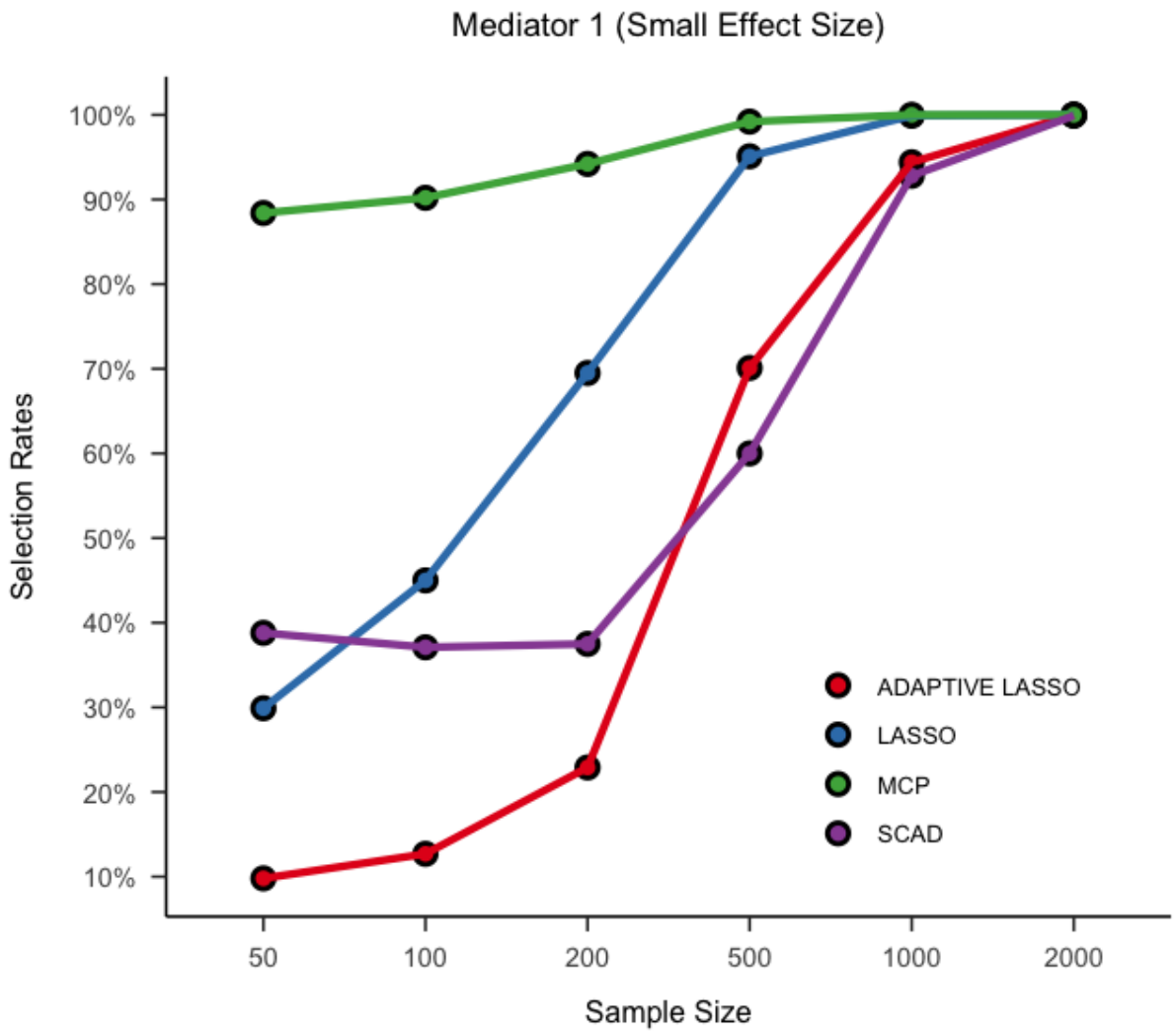

Figure 17. Line Plots comparing different regularization penalties for M_S. MCP had the highest selection rates whereas adaptive Lasso had the lowest selection rates for smaller sample sizes. All penalties levelled off at a sample size of 2000.

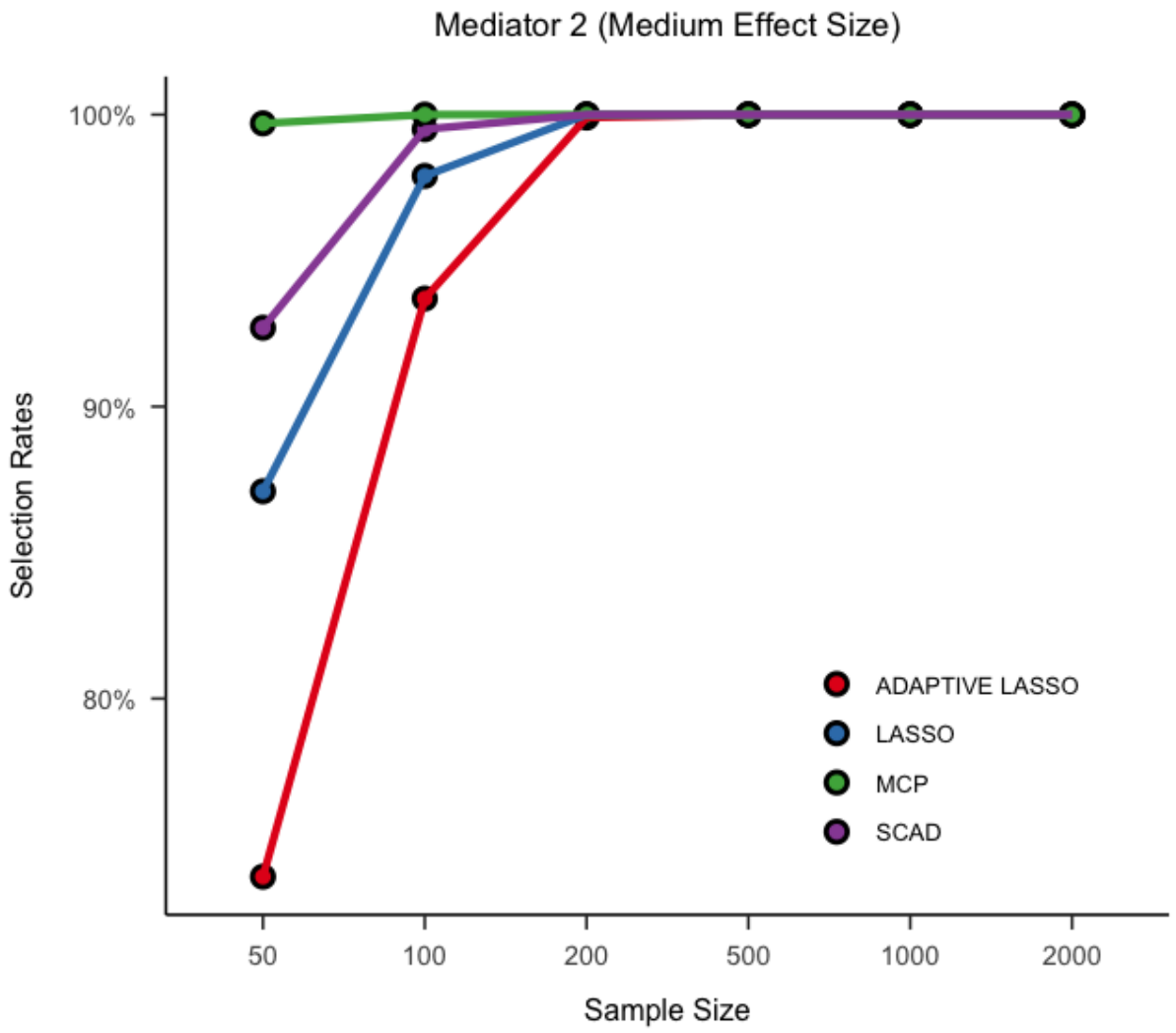

Figure 18. Line Plots comparing different regularization penalties for M_M. MCP had the highest selection rates whereas adaptive Lasso had the lowest selection rates for smaller sample sizes. All penalties levelled off at a sample size of 200. The scaling of the *Y* axis was changed to highlight minute differences.

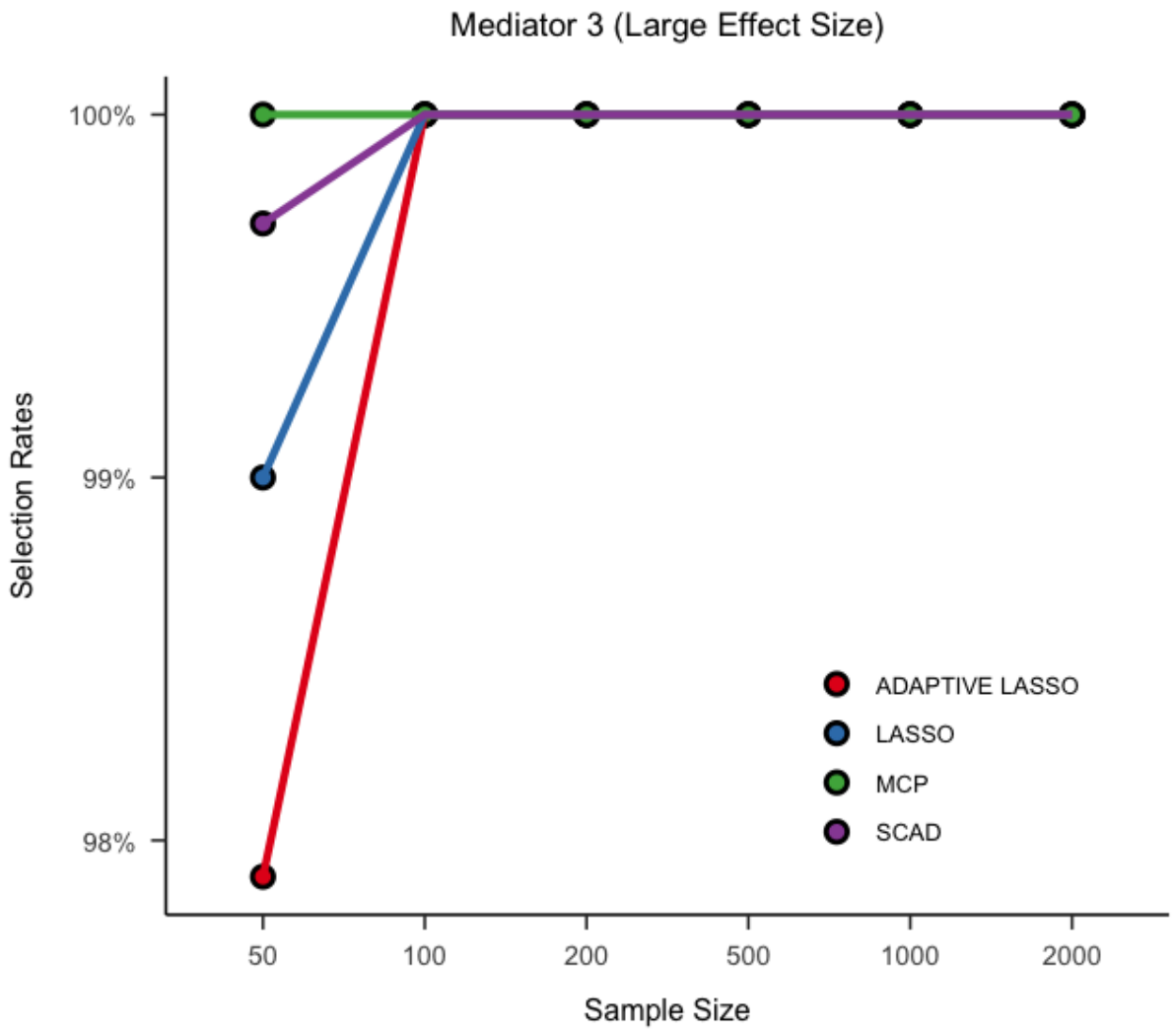

Figure 19. Line Plots comparing different regularization penalties for M_L. MCP had the highest selection rates whereas adaptive Lasso had the lowest selection rates for a sample size of 50. All penalties levelled off at a sample size of 100. The scaling of the *Y* axis was changed to highlight minute differences

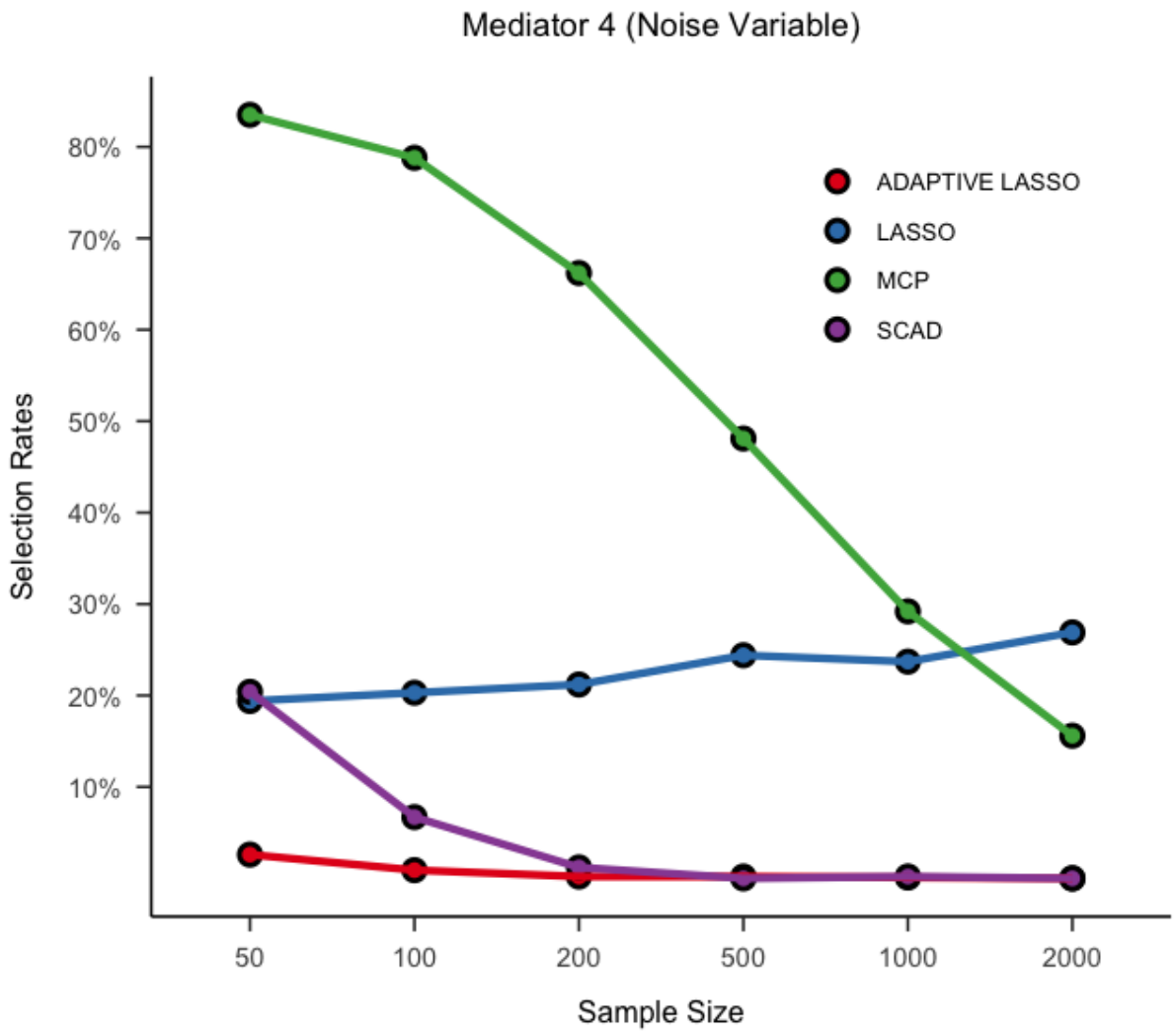

Figure 20. Line Plots comparing different regularization penalties for M_N. MCP had the highest Type I error rates whereas adaptive Lasso had the lowest Type I error rates. SCAD penalty began with moderate Type I error rates but approached 0 as sample size grew. The Lasso Type I error rate increased as sample size grew.

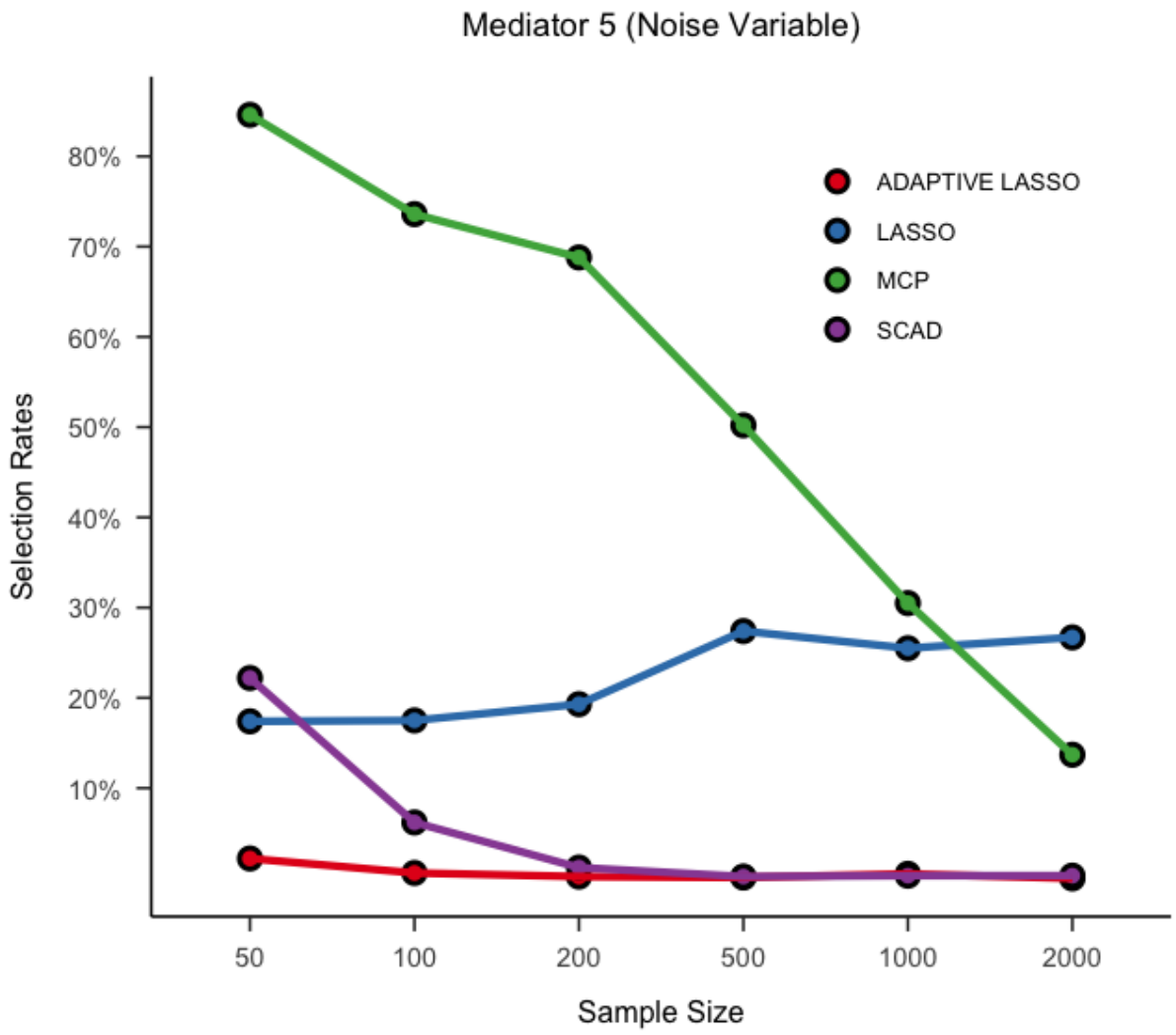

Figure 21. Line Plots comparing different regularization penalties for the second M_N. MCP had the highest Type I error rates whereas adaptive Lasso had the lowest Type I error rates. SCAD penalty began with moderate Type I error rates but approached 0 as sample size grew. The Lasso Type I error rate increased as sample size grew.

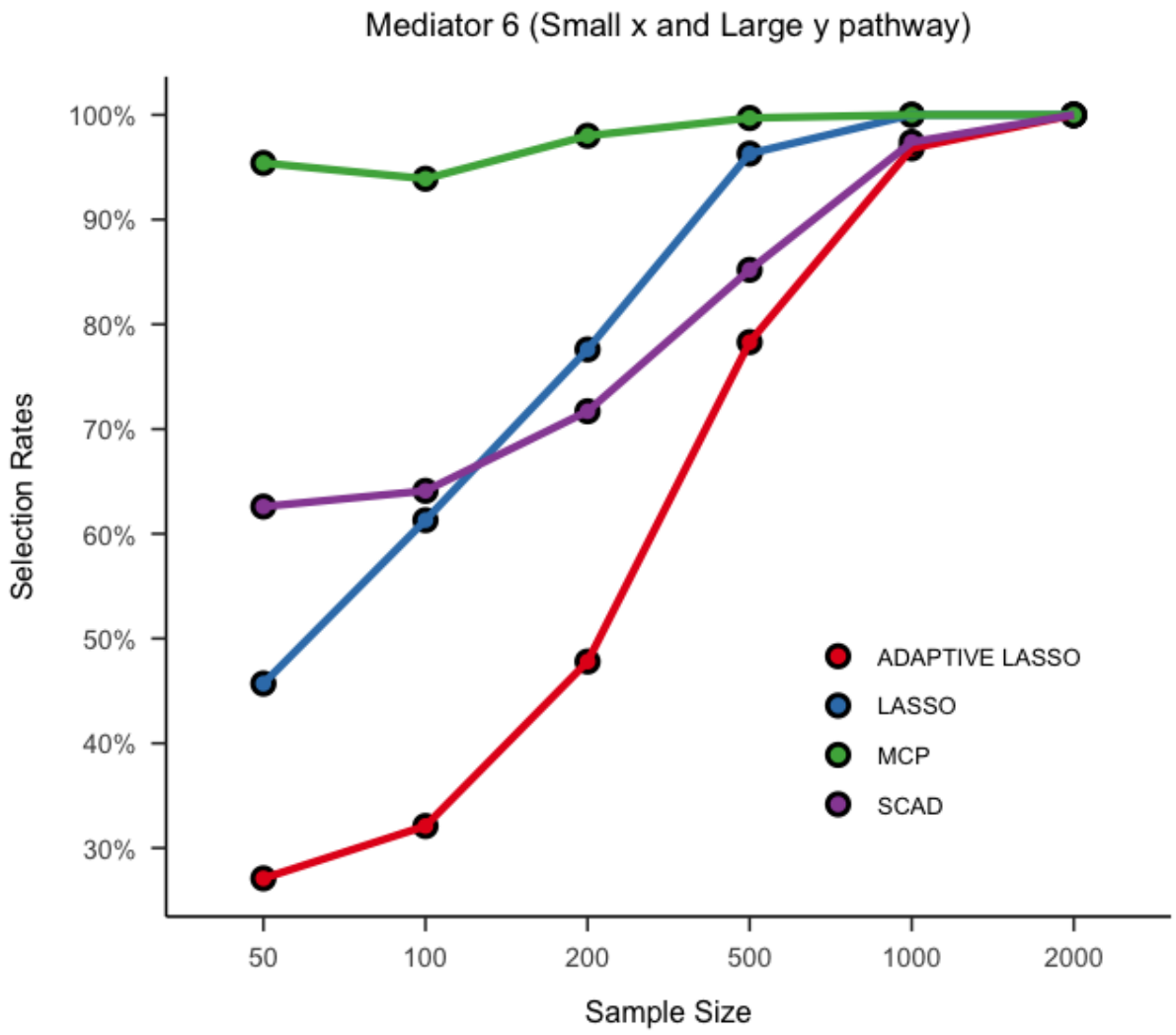

Figure 22. Line Plots comparing different regularization penalties for M_{SL}. MCP had the highest selection rates whereas adaptive Lasso had the lowest selection rates. SCAD penalty outperformed the Lasso for smaller sample sizes.

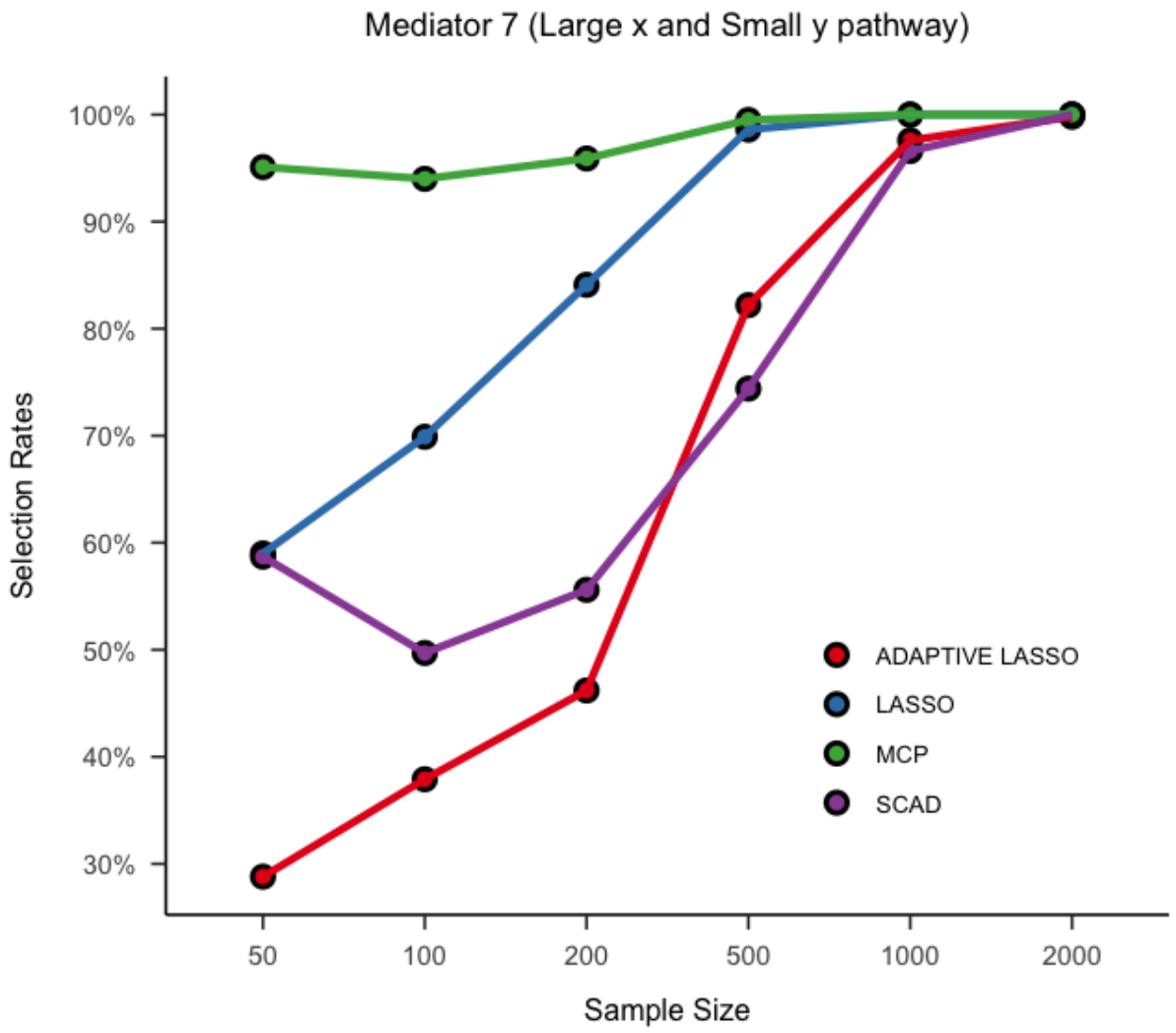

Figure 23. Line Plots comparing different regularization penalties for M_{LS}. MCP had the highest selection rates whereas adaptive Lasso had the lowest selection rates. The Lasso penalty outperformed the SCAD for all sample sizes.