The Importance of Random Effects in Variable Selection

A Case Study of Early Childhood Education

A THESIS PRESENTED
BY
THU PHAM
TO
THE DEPARTMENT OF STATISTICS AND DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
STATISTICS AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

APRIL 2023

Thesis advisor: Dr. Luke Miratrix and
Dr. Kelly McConville

Thu Pham

The Importance of Random Effects in Variable Selection

Abstract

Multilevel models are very common in statistical studies of education, as they account for the nested structure of many data sets. This paper combines this methodology with variable selection, motivated by the Early Learning Studies at Harvard (ELS@H). We use mixed-effects and fixed-effects LASSO in a simulation to gain insight into how random effects can impact variable selection. Furthermore, we also use the simulation to better understand the ELS@H data's weak relationship between the predictors and outcome.

From the simulation, we find that under settings of high cluster variance and high correlation between the clusters and covariate generation (which correspond to a low Intra-Class Correlation, or ICC), the fixed-effects and mixed-effects LASSO diverge in the variables they shrink. The ELS@H data shows a similar divergence, even though it exhibits a low ICC. We also find other general trends in how the fixed-effects and mixed-effects LASSO perform in various types of data. At the end of the paper, we are able to more strongly conclude that random effects are not contributing to the low predictive signal in the data. We also discuss whether considering random effects in variable selection is advantageous at large.

Contents

Listing of figures

To my parents and to my wonderful teachers, whom I've had the privilege of learning from.

Acknowledgments

To my two wonderful advisors, Luke Miratrix and Kelly McConville, thank you so much for sharing your helpful feedback, steady encouragement, and enormous wealth of knowledge. I am incredibly grateful for the opportunity to learn from your wisdom, as well as for the time you've generously given me during a busy year.

To Jonathan Seiden, Madelyn Garden, and the Early Learning Studies at Harvard, thank you for allowing me to join this study and providing sustained help throughout the process. I am inspired by your commitment to and expertise in early childhood education.

To Alex Young, Ariel Procaccia, Lucas Janson, and Joe Blitzstein, thank you for your invaluable mentorship and support for this thesis and my academic growth at large.

To my roommates and dear friends, thank you for getting me over the finish line and keeping me in good spirits. I am grateful for your patience, empathy, and company (and steady, generous surprises of caffeine and candy to keep me motivated) throughout this process.

Finally, to my family, I don't know where I would be without you all. I am grateful to my parents and grandma for their unwavering love and belief in me, and to my sister for her much-needed comic relief. This project is yours as much as it is mine.

1 Introduction

1.1 Motivation

A quality early childhood education (ECE) has been acknowledged by many experts as critical to youth development and success even in later years [1]. Despite this fairly universal agreement, there is little consensus on how to define and measure this "quality." Though (often non-causal) studies have attempted to resolve this lack of agreement, they have found little, and sometimes no, relationship between contemporary ECE measures of quality and childhood outcomes [2–6].

This thesis adds to this growing body of work by turning to LASSO variable selection. Although LASSO is not inference-based, it provides another statistical angle to approach ECE analyses. It has previously been used in education at large to improve predictive accuracy while

modeling student outcomes [7]. In this thesis, we use variable selection to filter which covariates are most predictive of early childhood outcomes in a specific dataset. Doing so takes a step towards identifying this previously mentioned elusive quality.

In this variable selection, it is important to account for the multilevel structure that often characterizes education data. Students are nested in classrooms, which are nested in schools, for example. This structure is important; students in the same classroom may have outcomes correlated with each other, due to their similar environments and interactions with each other, for example [8]. Failing to account for a multilevel data structure may lead to the misspecification of a model for the relationship between the predictors and outcome, which causes us to incorrectly conclude which covariates are related to the outcome. [8].

In summary, there is a very real need for further statistical investigation in the field of ECE. Variable selection can be used to sift through high-dimensional datasets, to narrow down which variables may truly be predictive of favorable early childhood outcomes. Because clustered data are so prevalent in education data, it is important to consider multilevel modeling in this variable selection context. This thesis seeks to better determine when this multilevel structure is important to consider in variable selection.

1.2 Problem

To address this question, we work with data from the Early Learning Study at Harvard (ELS@H), a representative longitudinal study of the ECE experiences of children in Massachusetts. The predictors in the ELS@H data convey a narrative of what is occurring in the classroom, through teacher and child observations. The outcomes are derived from well-established methods in ECE, which represent different

aspects of early childhood development. For example, the outcomes measure the children's ability to regulate their emotions, identify words, and solve basic applied problems.

However, upon a baseline linear regression (regressing the outcome of interest on all of the possible predictors), we see that there are only a few statistically significant predictors. This result, as well as the multilevel structure, makes us question how important the random effects component is. That is, is the random intercept contributing to the low predictive signal, or is there another reason?

This thesis attempts to answer this question for this specific data set through simulation. Our simulation compares the performance of two variable selection methods, one that considers random effects and one that does not. The purpose of our simulation is two-fold. First, we examine the general trends in the simulation under different settings related to how informative the covariates are and how extreme the clustering is. If we see agreement between the two variable selection methods, then we can more strongly conclude that the clustering in this setting does not affect the strength of the predictors. In addition, we compare certain simulation results to how the two variable selection methods perform in the actual ELS@H data. Specifically, if we see differing results in the simulation and the data, and we assume we have approximated the data well enough, then we can make more substantiated claims about the importance of clustering in the data.

1.3 Thesis Outline

This thesis begins by reviewing existing literature on multilevel models, LASSO as variable selection, and LASSO with random effects. Then, we move into an exploratory data analysis of the ELS@H data. We choose to gain insight into this data through simulation, which allows us to observe how varying the informativeness of the covariates

affects the comparison of fixed-effects and mixed-effects LASSO. We describe this multifactor simulation and run a targeted simulation that replicates the ELS@H data, as well as other simulations with different parameters. The results of the simulation are compared with the corresponding results in the ELS@H data, which can be used to address the original research question. Furthermore, some general trends in this type of multilevel data are explored.

2

Background and Methods

We use mixed-effects models and two types of LASSO, one that considers random effects and one that does not. Specifically, we repeatedly generate data with a mixed effects model and then fit both LASSO models to this data. The remainder of this section describes both the literature on these three methods, as well as the technical details behind them.

2.1 Mixed-Effects Models

Mixed-effects models are useful when there is some clustering in the data, or the observations exhibit a certain grouping structure. The membership in these different groups contribute to a relationship between the predictors and the outcomes that vary across

observations in different clusters. To model these varying relationships, mixed-effects models contain both random and fixed effects. In many statistical methods, we make the assumption that the observations are independently and identically distributed (i.i.d.). However, in the case of nested data, we can no longer make this assumption. Including random effects allows us to relax the i.i.d. assumption and to account for the clustering in the data.

To establish some terminology, mixed-effects models are also known as hierarchical linear modeling, multilevel modeling, mixed linear modeling, or growth-curve modeling [9]. In this thesis, mixed-effects models (or the random effects component of a mixed-effects model) and multilevel modeling will be the two terms we refer to. For simplicity, we assume that mixed-effects models refer to linear ones, which will be exclusively used in this thesis. In addition, mixed-effects models are a specific type of Generalized Linear Models and are sometimes referred to as Generalized Linear Mixed Models (GLMMs). Generalized Linear Models relax the assumption that the linear model must have a Normal error distribution [10]. Generalized Linear Mixed Models encompass this relaxation, as well as allowing random effects in the slopes, intercept, or both.

When adopting a fixed-effects model, we assume that there is a true treatment effect common across all observations. Any observed difference in effects is attributed to measurement error [11]. In contrast, random-effects models relaxes this assumption. Instead, observations in the same cluster are assumed to exhibit the same relationship between the predictors and outcome, but this is not necessarily true for observations across different clusters. There may be fundamental differences between observations (separate of the experiment) that cause them to have different treatment effects. Some examples of this clustering include workers nested in firms, patients nested in hospitals, and households nested in countries [12].

Mixed-effects models can encompass both random and fixed effects – they assume that there are some treatment effects that are constant across observations, while allowing other treatment effects to be random.

Moreover, using the proper degree of multilevel modeling is important for accurate prediction. Observations in the same cluster exhibit some interdependence. Intuitively, observations in the same cluster are more likely to have similar outcomes than observations in different clusters. This interdependence affects the variance of the outcome; if random effects are not accounted for, then we will underestimate the standard errors of the coefficients in a linear predictive model [13]. Incorrectly reducing the standard error gives us a narrower confidence interval of these estimates, lending false confidence in our estimates or associations. Finally, all of this mis-estimation leads to an increase in the possibility of making a Type I error, or concluding that a certain variable is statistically significant when it is not [14]. It is also detrimental to include unnecessary random effects – doing so will create a near singular random effect covariance matrix [15].

Multilevel modeling is applicable in many settings, such as public health, environmental studies, and sociology. [15–18]. In particular, multilevel modeling has been used in education. For example, if we are interested in a reading intervention's effect on student outcomes, it is possible that the treatment effect of the intervention on the student's reading scores may vary by classroom. This clustering can extend to more than one level; these classrooms are, in turn, nested in schools, and the schools are nested in school districts [9].

For that reason, multilevel modeling is a classic method in education statistical studies [9]. The i.i.d. assumption is rarely met when examining classroom performance. Students in the same classroom may exhibit some interdependence, through their

interactions with each other and the similar quality of instruction they receive from their shared teacher and curriculum. As explained earlier, this affects our statistical analyses of the students' performance.

In this thesis specifically, we apply mixed-effects models to nested data structures. That is, we can use these models when we know that our observations are clustered in different groups, which affects our analyses. We use different childcare providers to cluster the students; we have reason to believe that children in different schools experience different treatment effects. Thus, whenever we are interested in a specific child, it is necessary to include information that identifies their early childcare provider (denoted j).

Another alternative is to instead regress on the clustering variable. However, we note that there are many clusters present in our data (specifically, over 150; see Table 3.4.1), which would require many different covariates to account for all the levels in the categorical variable that denotes the cluster. There are already many predictors in the data set, which makes adding an additional 150+ predictors untenable. Furthermore, multilevel modeling is such a classic method in education that it would be remiss to ignore its use in the context of early childhood outcomes.

2.1.1 Technical Details

Statistically, mixed-effects models are implemented by allowing the intercept, slopes, or both to be drawn from a random distribution, instead of fixed at a constant. It is possible for a model to have both random intercepts and random slopes. However, this thesis only uses random intercepts, mostly in the interest of computational capacity and to exercise extra caution in including too many random effects. Thus, with a random intercept, y_{ij}, which is the outcome of the ith observation in the jth cluster, can be expressed as a function of the

following [18, 19]:

$$
\begin{aligned}
y_{ij} &= b_j + \mathbf{x_{ij}}\boldsymbol{\beta} + \epsilon_{ij} && (2.1)\\
b_j &= \gamma + U_j && (2.2)\\
\mathbf{x_{ij}} &= \begin{bmatrix} 1 & x_{1,ij} & \dots & x_{p,ij} \end{bmatrix}\\
\boldsymbol{\beta}^T &= \begin{bmatrix} \beta_0 & \beta_1 & \dots & \beta_p \end{bmatrix}\\
\epsilon_{ij} &\sim \mathcal{N}(0, \sigma_n^2), \text{ where}
\end{aligned}
$$

$\boldsymbol{\beta}$ are fixed effects across clusters, which holds true for all p predictors. Because they are fixed, it is unnecessary to include the j in their subscripts. In this format, the fixed intercept is absorbed into these fixed effects, and we ensure it is properly included with an additional 1 in the vector of covariates. σ_n^2 is the variance for the Gaussian random noise.

On the other hand, b_j varies with the cluster that the observation is in and can be decomposed into the mean intercept for all J clusters, γ, and the unique effect of cluster j on an intercept, U_j (Equation 2.2). U_j is often generated from a Normal distribution $\mathcal{N}(0, \sigma_r^2)$ and variance σ_r^2 [18].

We can further express the outcome of the observations within the jth cluster [19].

$$\mathbf{Y}_j = \mathbf{b}_j + \mathbf{X_j}\boldsymbol{\beta} + \boldsymbol{\epsilon}_j \tag{2.3}$$

$\mathbf{b}_j$ has a single unique element (the random intercept for the jth cluster) that is repeated across n_j rows, where n_j is the number of observations in the jth cluster. $\mathbf{Y}_j, \mathbf{X_j}$ are the outcomes and design matrices, respectively, for all of the observations in cluster j. Finally,

$\boldsymbol{\epsilon}_j$ is a vector of length n_j of draws from the random noise distribution.

Finally, we can express the outcome for any observation, regardless of cluster [19]:

$$\mathbf{Y} = \mathbf{b} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.4}$$

A metric often used in the context of multilevel models and clustering is the Intraclass Correlation Coefficient, or the ICC. The ICC is the proportion of the total variance explained by the clustering structure [20]. We compute this metric in each simulation iteration and compare it with the ICC in the ELS@H data.

2.2 LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) Regression is a type of linear regression method. Like any other prediction modeling, it seeks to maximize predictive accuracy. It does so by choosing a model that minimizes a loss function, with an added penalty for the sum of the absolute magnitude of the coefficients. Thus, LASSO optimizes for predictive accuracy and also seeks to avoid overfitting.

LASSO is especially useful in high-dimensional settings; that is, when $p \gg n$, where n is the number of observations and p is the number of predictors. LASSO often outperforms over other high-dimensional methods in computational ease, as it only relies on convex optimization [21]. Finally, LASSO has proven superior to standard maximum likelihood estimators in minimizing the Type I error rate in certain contexts [22–25].

The most recent LASSO literature has been concerned with optimizations. Researchers have especially focused on improving its

computational efficiency, as well as achieving the oracle property, which occurs if the LASSO performs as if it had received the underlying model in advance [26].

2.2.1 Technical Details

To optimize prediction accuracy while also avoiding overfitting, LASSO maximizes the likelihood with a constraint. This method is also equivalent to maximizing the log-likelihood, minimizing the negative log-likelihood, or minimizing the loss function. These are all possible so long as there is a constraint that the sum of the absolute values of the coefficients is less than or equal to some pre-determined quantity. In this project, we minimize the loss function. This form of LASSO operates on an underlying model with fixed effects only (not the same as equation 2.1), where

$$y_i = \beta_0 + \mathbf{x_i}\boldsymbol{\beta} \tag{2.5}$$

Thus, $\hat{\boldsymbol{\beta}}^T = \begin{bmatrix} \beta_1 & \dots & \beta_p \end{bmatrix}$ can be derived from the following minimization problem [23]:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1} \left(y_i - \beta_0 - \mathbf{x_i}\boldsymbol{\beta}\right)^2, \\ &\text{subject to} \sum_{j=1} |\beta_j| \leq t \qquad\qquad 0 < t < \infty \end{aligned} \tag{2.6}$$

Alternatively, we can write this minimization problem in the Lagrangian form, with a shrinkage penalty $0 < \lambda < \infty$ [23].

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1} \left(y_i - \mathbf{x_i}\boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.7}$$

With this minimization problem, all coefficients of covariates are shrunk toward zero with some coefficients being set to zero. This feature contrasts with that of Ridge regression, LASSO's closely related counterpart, where coefficients are shrunk towards, but never to, zero [23]. This thesis uses LASSO as opposed to Ridge for that reason, since covariates are more clearly "selected."

The regularization hyperparameter λ needs to be tuned before fitting the final LASSO model.

2.2.2 Tuning The Regularization Parameter

The shrinkage penalty λ is a hyperparameter that needs to be tuned before the LASSO regression is fit. Five-fold cross-validation is performed with a range of finite, positive values $\lambda_1, \ldots, \lambda_L$, where $L = 100$ to determine which λ_i value is best for the data. For each λ_ℓ the train data is split into five equally-sized subsections; four of those are used to fit a model with the λ_ℓ value, and the last fold is used to compute a measure of how well that model fits the data. We choose a criteria for this measure ahead of time, and the λ_ℓ with the minimum chosen criteria is chosen as the optimal value of λ. There is a built-in, rather efficient, cross-validation feature of the LASSO in R called `cv.glmnet`, which we take advantage of. The default settings of `cv.glmnet` are an L value of 100 and a criteria of mean squared error.

2.2.3 LASSO as Variable Selection

Besides creating a more predictive model, LASSO can also be used for variable selection. Any variables that have a non-zero coefficient in the final model are deemed as "predictive" of the outcome of interest. Notably, it has been adapted for variable selection in Cox's proportional hazards model, a type of regression model commonly used in statistical medical research [27]. This same paper found that,

in this application, using LASSO was more accurate than forward stepwise selection, a more explicit and conventional form of variable selection. Furthermore, LASSO has been used for variable selection in genetic data, where there is often little prior information about which variables are most informative [28]. In general, LASSO is attractive because it addresses both overfitting and the overestimation of how well a model performs (in the case of including all possible variables).

Using LASSO for variable selection is a statistically valid choice. To substantiate this claim, previous literature has assumed a Multivariate Normal distribution in data and represented any conditional independence relations through graphs.[1] As usual, the goal is to estimate the true model, which can either be attained by estimating the conditional independence relations (also known as covariance selection), or through standard variable selection [29]. LASSO can perform neighborhood selection, separately estimating the conditional independence relations by optimizing a convex function for each of the p nodes (for p covariates). This method is much more attractive than methods that consider all possible models (a quantity on the order of p^2); thus, these results substantiate the ongoing theme that LASSO is generally more computationally efficient when estimating the true model among high dimensional, sparse data [30]. The authors of this same paper found that the probability of using LASSO to estimate the true neighborhood (and thus, the true model) converges to one.

We intentionally choose LASSO over subset selection, forward and backward stepwise regression, and Ridge regression. Subset selection retains a certain subset of predictors in the final linear model. Discarding less predictive covariates can be favorable – the model becomes more interpretable and improves its ability to predict

[1] To generate a conditional independence relation graph from Multivariate Normal data, an edge is present between nodes a and b if and only if X_a and X_b are conditionally dependent.

out-of-sample. However, subset selection, because it entirely retains or discards variables, is subject to high variance. Furthermore, it is computationally intensive [26]. A workaround lies in greedy algorithms, but these are vulnerable to getting "stuck" at a local optimum. There have been substantial improvements that make these algorithms more accurate, but the question of computational intensity would still be relevant [31]. On the other hand, LASSO *shrinks* the Ordinary Least Squares estimators, a more continuous process. Thus, LASSO is not as vulnerable to high variance, and can more consistently reduce prediction error as compared to the model that includes all variables [23].

As mentioned above, LASSO more clearly selects variables than Ridge because it shrinks certain coefficients to zero. A previous simulation study has shown that LASSO outperforms best subset regression in low Signal-to-Noise Ratio settings, which is more representative of the ELS@H data (the data we are trying to emulate) [32]. In the same simulation study, the authors state that the forward and backwards stepwise regression are seen as a local optimum of the best subset regression [32]. Given all of this information about other variable selection methods, this project opts to use LASSO.

2.3 LASSO with Random Effects

It is widely agreed that mixed effects are important to model many contexts, but unfortunately, these models can become unstable in the presence of many covariates [19]. This problem highlights the importance of variable selection in multilevel data. The LASSO described above only accounts for fixed effects, introducing the need for a LASSO that also accounts for random effects. As described in the previous section, this inclusion of random effects means we have a lower probability of misspecifying the model.

Variable selection while considering random effects is a fairly new and generally unexplored area of research. A lot of hesitance around further research in this field relates to the computational difficulty; because both random and fixed effects are now considered, it is possible for the number of possible models to grow exponentially with the number of predictors [15]. Despite this uncertainty, LASSO remains a favorable variable selection for random-effects models. There are other methods that avoid searching through the entire model space, but they are still computationally intensive when there are many predictors [15]. The existing research around LASSO justifies its existence by comparing it to other prediction models and speculates on the order of selecting effects.

To begin this literature, a 2010 paper introduced an $\ell 1$-penalized estimation procedure (which is the same penalization in Equation 2.7) for high-dimensional multilevel data. They mathematically proved that this procedure possesses both consistency and oracle properties. Finally, they also used a simulation and a biology data set to show that including the random effects in LASSO can reduce the variance in the predictive error [21].

A simulation carried out by W. Holmes Finch in 2018 has proven that a Multilevel LASSO (MLL) often outperforms the standard restricted maximum likelihood estimator (REML) in high dimensional multilevel data [18]. Finch was primarily concerned with whether the 95% confidence intervals for the two methods captured the true coefficients and whether they correctly detected a non-zero variable in the model (relating to the Type I error and power). The two were comparable in parameter estimation bias and standard error. However, MLL yielded higher coverage, a lower Type I error, and higher power [18]. Thus, Finch showed support for using the Multilevel LASSO instead of the standard REML method. In other words, the simulation proved that the regularization of data with

random effects is also important and can improve outcomes.

There has also been research to demystify how exactly to select for random effects, since an entire row and column of the covariance matrix must be eliminated for a random effect to be shrunk to zero. Bondell et al. argues for the simultaneous selection of random and fixed effects, as selecting the pertinent random effects before the fixed effects may yield different results from the reverse order of selection [15]. In a simulation and data study, they compared the Restricted Information Criterion (REML.IC), which performs selection on the variance-covariance structure before selecting fixed effects; the Extended Generalized Information Criterion (EGIC), which selects the fixed effects (while including all of the random effects in the model) before selecting the random effects; and the simultaneous selection they propose. The authors found that their simultaneous selection method correctly identified the true model most often, and the parameters chosen by the simultaneous method correspond to a higher likelihood value. Specifically, they emphasized that the sequential selection models often select effects incorrectly in the first step, which lowers the accuracy of the second step of selection. To further advocate for their method, they proved that their proposed likelihood estimator identifies the true model with probability tending to one [15].

We contribute to this rather nascent area of literature by comparing the mixed-effects and fixed-effects LASSO in settings that have not been explored in previous simulations. Furthermore, we use mixed-effects LASSO in an area (education) where it had not been previously introduced.

2.3.1 Technical Details

As mentioned above, this project only uses a random intercept as part of the random effects. Thus, it is unnecessary to shrink any random effects that correspond to slopes. Below, we describe the corresponding mixed-effects LASSO, which only shrinks the fixed effects while also considering the random intercept.

To do so, the log-likelihood includes random effects and penalizes the fixed ones [19]. We let $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{b})$ be the set of parameters in a multilevel model, according to the specification in equation 2.4.

We assume that the distribution of the random intercepts $\boldsymbol{b}$ follows a Multivariate Normal, whose covariance matrix is $\boldsymbol{Q}$. We use $\boldsymbol{\gamma}$ to denote the unknown parameters that determine the covariance matrix of the random effects, $\boldsymbol{Q}$.

The likelihood, or conditional density, is conditioned on both the covariates and the random effects, and is denoted for the ith observation in cluster j, y_{ij}, as $f(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_j)$, for $j = 1, \ldots, J$ and $i = 1, \ldots n_j$. The larger vector $\mathbf{b}$ is the concatenation of the random intercepts for all J clusters.

To represent all observations in a cluster j, we use the vector notation $\boldsymbol{y}_j$. With the correctly specified parameters $(\boldsymbol{\delta}, \boldsymbol{\gamma})$, $f(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_j)$ should be equivalent to $f(\boldsymbol{y}_j \mid \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{b}_j)$. To get to our desired likelihood, $\ell_j(\boldsymbol{\delta}, \boldsymbol{\gamma})$ for the jth cluster, $\boldsymbol{b}_j$ must be marginalized out. Using the continuous Law of Total Probability and $p(\boldsymbol{b}_j, \boldsymbol{\gamma})$ for the density of $\boldsymbol{b}_j$:

$$f(\boldsymbol{y}_j \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) = \int f(\boldsymbol{y}_j \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) p(\boldsymbol{b}_j, \boldsymbol{\gamma}) d\boldsymbol{b}_j \tag{2.8}$$

Finally, to achieve the complete likelihood, we must use all observations across J clusters [19]:

$$\ell(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \log\left(\prod_{j=1}^{J} \int f(\boldsymbol{y}_j \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) p(\boldsymbol{b}_j, \boldsymbol{\gamma}) d\boldsymbol{b}_j\right)$$
$$= \sum_{j=1}^{J} \log\left(\int f(\boldsymbol{y}_j \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) p(\boldsymbol{b}_j, \boldsymbol{\gamma}) d\boldsymbol{b}_j\right) \tag{2.9}$$

Estimating the inner term calls for the Laplace method, which is typically used to estimate integrals of exponential expressions. This method yields the following likelihood for a generalized linear mixed model, which was derived by Breslow and Clayton [33].

$$\ell^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \left(\sum_{j=1}^{J} \log(f(\boldsymbol{y}_i \mid \boldsymbol{\delta}, \boldsymbol{\gamma}))\right) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{Q}^{-1} \boldsymbol{b} \tag{2.10}$$

Finally, with all of this notation, we can define a penalized likelihood. The penalized likelihood is then maximized to solve for the estimated coefficients of the fixed effects. Because this maximization problem is rather unwieldy, a 2011 paper proposed a gradient ascent algorithm to estimate the coefficient values [19].

$$\ell^{\text{pen}} = \ell^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i| \tag{2.11}$$
$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} [\ell^{\text{pen}}] \tag{2.12}$$

2.3.2 Choosing The Regularization Parameter

Similar to the selection of λ for a fixed-effects LASSO, we use cross-validation as described in the earlier subsection. However, because there is no built-in cross-validation function for the

fixed-effects LASSO, a few design decisions needed to be made. To determine a range of regularization parameters to cross-validate from, λ_{min} is the largest λ value for which all of the covariates are retained, and λ_{max} is the smallest λ value for which all of the covariates are shrunk to zero. We then proceed with the five-fold cross-validation.

We make one key modification to account for the computational intensity of the mixed-effects LASSO. Between simulations generated with the same Signal-to-Noise ratio, the λ_{min} and λ_{max} values do not differ drastically. Thus, before running the full simulation, we find the average λ_{min} and λ_{max} values across twenty repetitions, considering twenty λ values each time. In this way, for each simulation run, we do not need to recompute a new λ range, with little deviation from the true values of λ_{min} and λ_{max}. This modification saves us time during a rather computationally intensive simulation.

3 Data

3.1 Data Source

Our data was provided by the Early Learning Study at Harvard (ELS@H), a longitudinal study of the ECE experiences of children in Massachusetts. Children in the study are representative of the ethnicities, home languages, incomes, and other demographics of families in Massachusetts. The different settings of preschools are also represented in the ELS@H data. The four year study began in 2017, beginning with around 3,500 3-4 year olds across the state. The study followed these children and their families until they aged out of preschool (which is usually at age five) [34]. For sake of data availability, we use years one and two from the study, 2017-18 and 2018-19.

3.2 Data Description

The predictors in the ELS@H data come from the Child Observation and Teacher Observation in Preschools (COP-TOP), which constructs a narrative of what is happening in each classroom through a rich set of variables. Both of these measures were originally developed by the Peabody Research Institute at Vanderbilt University, and adapted for ELS@H's purposes. The COP-TOP variables have a longitudinal aspect to them, which comes from the multiple sweeps of children and teacher observations throughout the years.

The intent of the Child Observation in Preschools (COP) component is to understand how the child spends their time in a classroom as a whole, and to compare children's different experiences [35]. The Teacher Observation in Preschools (TOP) is very similar, painting an overall picture of a teachers' actions towards their classroom [36]. At the individual child level, the COP and TOP predictors are primarily categorical variables. For example, a specific COP variable is "type of task," which has various levels for different activities that the child could be engaging in at the time of that sweep.

The child-level outcomes in the data are derived from other studies proven to represent several domains in early childhood development. The outcome that we focus on is the difference in these child-level outcomes from year one to year two. Furthermore, since these outcomes are measured within the environment of the students' providers, we cluster the students by the different providers. These providers are representative of childcare in Massachusetts and include public preschool, Head Start [1], family child care, and community centers.

We only use the COP-TOP variables from year one (2017-18) and

[1]Head Start is a federal program that provides funding for low-income children to enroll in high-quality ECE programs.

the difference in child-level outcomes from year one to year two. Essentially, we are interested in the environment of the children's classrooms in year one, and their change in outcomes after that year in the environment.

3.3 Data Cleaning

We begin with a data set that consists of the COP-TOP predictors in year 1 and the growth in outcomes from year 1 to year 2 for children who were observed at least ten times.[2] This particular data set has 1,163 observations (where each observation is a child), 9 outcomes, and 89 predictors. Many of the original variables are categorical, with three or more categories. Any purely quantitative variables were standardized to have a mean of zero and a standard deviation of one. To make computations of these categorical variables meaningful, we expand these categorical variables into dummy variables. Furthermore, as a proxy for the classroom environment, we compute the leave-out mean and leave-out standard deviation for each child's set of predictors, based on the classroom they are in. That is, for each child and each predictor p_i, we compute the classroom mean and standard deviation of p_i, excluding that particular child.

In addition, before choosing an outcome and doing some pre-processing of predictors, we did not use the original longitudinal nature of the data. Instead, we averaged the values taken at each sweep for an individual child. Finally, we also discarded covariates that uniquely identify a child, since now each row represents a child.

After expanding the categorical variables into dummy columns,

[2] In a preliminary study of variable selection with this data set, any children who had were observed less than ten times were dropped. We maintained a lot of the data cleaning practices from the previous study, which explains the reduction in observations from the original data set. The conference study for this previous study can be viewed here.

averaging the children's covariates across sweeps, and computing the leave-out mean and standard deviation, there are 1,663 unique children, 278 predictors (excluding the clustering variable), and 9 outcomes. There are three different cleaning processes, which are described in detail below and result in different dimensions of the data.

3.3.1 Choosing an Outcome

There are nine total outcome variables. We note that our main goal is to understand whether the noise in the data is due to its multilevel structure, or the fact that the predictors are truly weak. This goal is not necessarily furthered by considering multiple outcomes, which would require a method like the multivariate group sparse LASSO [37]. Furthermore, the simulation of the data only involves creating one outcome; thus, we also need to clean the data to select an outcome of interest.

We see that the nine outcomes exhibit some pairwise correlation with each other (through a correlation matrix), and so we choose a few different outcomes based on different criteria. First, we choose two outcomes that are most "representative" of all the outcomes, based on two different perspectives.

The first is choosing the outcome with the highest multiple correlation with the other outcomes. We find this outcome by computing the precision matrix of all eight outcomes, which is the inverse of the correlation matrix. The diagonal elements of the precision matrix are proportional to the conditional variances of that outcome given all other outcomes [38]. The conditional variance is also known as multiple correlation, or, in this context, how well that outcome can be predicted from all the other outcomes. We choose the outcome with the maximum multiple correlation (i.e., the diagonal

element of the precision matrix with the highest value), which is `gain_c_ltr_cogsoc_comp`. For an individual child, this outcome is the cognitive and social score derived from their performance on the Leiter Test. [3]

Another method is to create a new outcome using Principal Components Analysis (PCA). PCA is typically used for dimension reduction, but we extract the first principal component (PC1), which is a vector formed from the nine different outcomes that best captures their variance. Thus, we form a hybrid outcome from the linear combination of all nine outcomes given by the PC1. One drawback to this method is that we chose to exclude observations with an NA value for *any* of the outcomes had to be excluded, in line with ELS@H's prior treatment of NA values. There are few children for whom all nine outcomes were recorded, so the data set associated with this hybrid outcome is also small.

The final alternative is to explore the outcome that is the "least" representative of all other outcomes. To accomplish this, we look at the outcome with the lowest multiple correlation with all other outcomes, which is `gain_c_mefs_str`, the child's Z-score relative to the national average of the Minnesota Executive Function Scale.[4]

3.3.2 Pre-Processing of Predictors

Before any pre-selection of predictors, there are a total of 278 predictors. This amount of predictors makes mixed-effects variable selection computationally intensive, as the existing package is relatively new and not as efficient as the standard fixed-effects LASSO. Furthermore, this mixed-effects LASSO is not equipped to

[3] The Leiter Test assesses cognitive functions in children and adolescents through nonverbal tasks, mainly requiring matching or pointing responses [39].

[4] The Minnesota Executive Function Scale provides a direct behavioral measure of Executive Function skills, or the child's ability to learn and execute certain tasks, as well as regulate various behaviors. [40]

handle perfectly collinear or even highly correlated covariates, unlike its fixed-effects counterpart. In particular, when these highly correlated or perfectly collinear covariates are present, there are errors with model matrices that do not have full rank, or a non-invertible Fisher matrix. These errors occur when there exist linear combinations among the predictors, or the predictors are highly correlated enough that a linear combination of some predictors can closely approximate another predictor. Thus, only a subset of covariates was picked for the analysis, to avoid any fatal errors.

The pre-processing involved four total steps. The first step was using fixed-effects LASSO with a low regularization parameter of 0.01, which brought the number of predictors to 220. Next, any covariates that had a high frequency of high pairwise correlation (i.e., over 0.95) were eliminated. Specifically, we first record the frequency of each predictor as it appears in a pairwise correlation of greater than 0.95.

There is then an iterative removal process, which is also explained in pseudocode in Algorithm 1. First, we remove the predictor that has the highest frequency of appearing in a pairwise correlation greater than 0.95. Each time a predictor is removed, the aforementioned frequency "table" is updated. For example, when we remove a predictor p_i, then any other predictors $p_j, j \neq i$ that had a pairwise correlation with p_i greater than 0.95 will also appear one less time in the counts of high correlation pairs. Thus, the frequency with which other predictors p_j in the pair (p_i, p_j) appear must be updated each time we remove a variable p_i. This process repeats until the frequency of predictors that have this high pairwise correlation is at most 8. The final step in this pre-processing step is to remove any variables that are a linear combination of each other. Please refer to Table 3.4.1 for the final dimensions of these different data sets.

Algorithm 1 Pseudocode for iteratively removing covariates that appeared in pairs with high correlation. P stores each pair of covariates X_i, X_j and their corresponding correlation c. Furthermore, the function count(P, t) finds the number of covariate pairs that have a correlation larger than a threshold t and the function remove(X_r, P) removes any pairs containing X_r from the table P. Finally, the function find_max(P) finds the X_i in (X_i, X_j) that corresponds to the highest correlation in P.

$n \leftarrow 8$
$i \leftarrow \text{count}(P, t)$
$\ell \leftarrow [\,]$
while $i > n$ **do**
    $X_r \leftarrow \text{find\_max}(P)$
    $\ell \leftarrow [\ell; X_r]$
    $P \leftarrow \text{remove}(X_r, P)$
    $i \leftarrow \text{count}(P, t)$
end while
return ℓ

3.4 Exploratory Data Analysis

Finally, it is necessary to investigate certain quantitative characteristics of the ELS@H data set, in order to inform the simulation. Unfortunately, some aspects of the data cannot be captured with the parameters that determine the simulation. However, we can choose parameters in the simulation that are as close as possible to the data. We explore the data from the perspective of the three aforementioned outcomes, since the parameters are slightly different under each outcome.

First, to estimate the underlying coefficients of the data, we run a linear regression on all variables but the clustering coefficient (i.e., the ID of the provider).

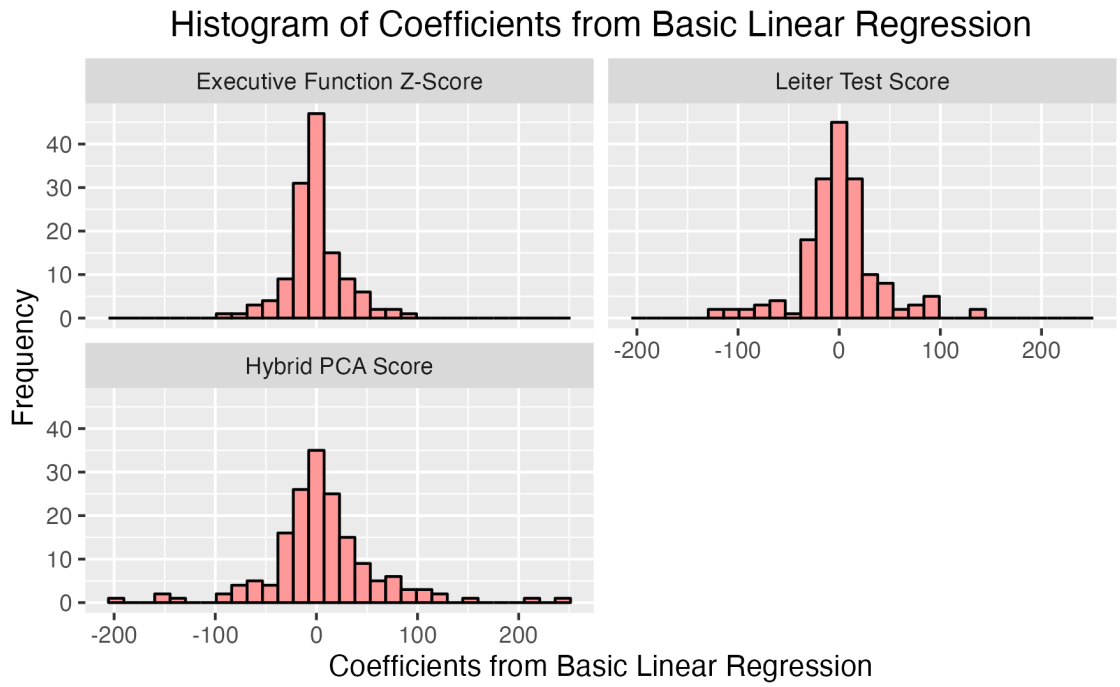

Figure 3.4.1: A faceted histogram of the value of the coefficients estimated by an all-inclusive linear regression, using all three outcomes.

The important takeaways from figure 3.4.1 is that there are both positive and negative coefficients, and that the highest frequency of coefficients is clustered around a coefficient value of zero. There are also several outlier coefficient values.

Another useful distribution to examine is that of the cluster sizes.

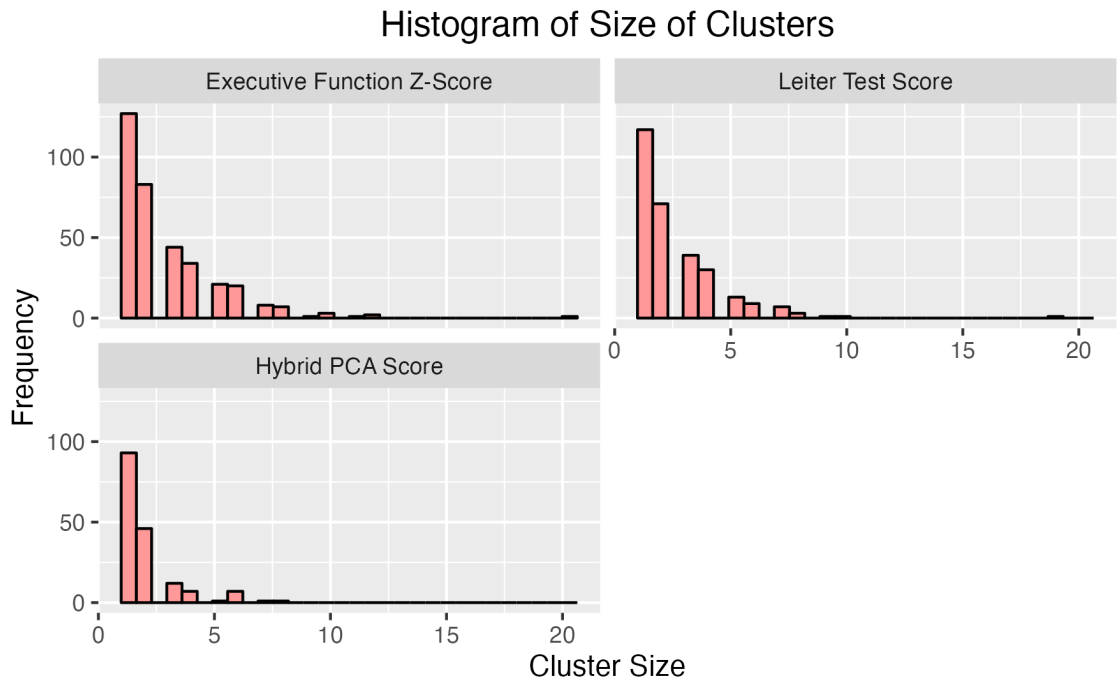

Figure 3.4.2: A faceted histogram of the cluster sizes present in the data, for all three outcomes. The hybrid PCA score has fewer clusters because of the data loss (described above), which was required to compute the hybrid PCA outcome.

The sizes of the clusters are mostly fewer than five students, with clusters as small as one student; a good number of these are smaller providers, such as family childcare. Similar to the examination of the coefficient values, there are a few clusters with extreme values (i.e., ten or more students).

Other important characteristics of the data are listed in Table 3.4.1. One important note is that for the Leiter Test, the multilevel LASSO had issues with invertible Fisher matrices when given a dataset that only excluded predictors with pairwise correlation of greater than 0.95 (the threshold used for the other two outcomes). Instead, a threshold of 0.80 for the pairwise correlation was used. As a result, there are noticeably fewer predictors in the data corresponding to Leiter Test

outcome.

Parameter	Description	MEFS Value	Leiter Value	PCA Value
J	The number of clusters, or providers, in the data	232	352	168
$\bar{n}$	The average size of each cluster	2.49	2.82	1.85
var($\bar{n}$)	The variance in size of each cluster	3.96	5.35	1.86
snr	Signal-to-noise ratio	0.42	0.23	1.28
p	Number of predictors, excluding the clustering variable (provider ID)	173	133	169
cov(X)	Average covariance between non-clustering covariates	0.08	0.04	-0.006
icc	Intra-class correlation	0.01	0.13	0
w	Number of coefficients with an absolute value larger than 20 from a linear regression that includes all predictors	71	45	85
N	Total number of unique children represented in data	726	992	311
R^2	Proportion of variance explained by a basic linear model	0.29	0.18	0.58

Table 3.4.1: Important characteristics of the data with outcomes `gain_c_mefs_str`, `gain_c_ltr_cogsoc_comp`, and the PCA-derived one (from left to right).

In particular, we justify the determination of the w parameter. The maximum absolute value of these coefficients is around 200. Thus, in a way, we set the threshold that determines whether a predictor is

"strong," by dividing this maximum by 10. This value of w is one of the parameters of the simulation. Although the determination of this threshold was somewhat arbitrary, it was a decision that was important to move forward with the simulation.

4 Simulation

4.1 Overview

The purpose of this simulation is to explore the importance of random effects in variable selection. One way to answer this question is to investigate whether considering mixed effects affects the variables that are selected, in different settings of data. For prediction purposes, we are also interested in various error measures relating to the predicted outcomes and coefficients.

In our simulation, we generate multilevel data that differ on the informativeness of the covariates, the variance in random intercepts, and the degree to which the clustering affects the generation of the covariates. We then use the `glmmLasso` and `glmnet` LASSO packages to select the most predictive covariates considering both mixed and

fixed effects, respectively. Finally, we evaluate the quality of these models by several metrics, specifically those surrounding each model's ability to select the covariates that are predictive of the outcome.

Ultimately, we use this simulation to better understand the ELS@H data. The main parameters of interest (i.e., the ones we vary in the simulation) are those relating to how informative the predictors are and how extreme the clustering is. As previously mentioned, the data already has a very low adjusted R^2 value, and only a few predictors that are flagged as statistically significant. We are primarily interested in showing how the fixed-effects LASSO and mixed-effects LASSO differ in the variables they select, in simulations with different settings of covariate informativeness and cluster-related parameters. In general, a higher agreement indicates that the multilevel structure of the data is not as important to the strength of the predictors (in that particular setting). Examining trends of agreement between the fixed-effects and mixed-effects LASSO in these different settings can help us understand in what situations the two methods should agree.

We simulate some data that very closely mimics the ELS@H data, and compare the agreement in the simulation and that of the ELS@H data. We also examine the general relationship between certain parameters and the agreement of the two methods, as well as other performance metrics of interest. We would also like to credit Pustejovsky and Miratrix for their Monte Carlo Simulations guide, which provides a lot of the basis code for this simulation [41].

4.2 Setup

This project simulates normally distributed two-level data and compares two models: one that considers both random and fixed effects, and one that only considers fixed effects. The true data-generating model is one with random intercepts, mimicking the

random effects that students across different providers may experience. Data is simulated from the following model, where $J_{r,c}$ is a $r \times c$ matrix of ones:

$$Y'_{ij} = \pi_j + \mathbf{X_{ij}}\boldsymbol{\beta} \tag{4.1}$$

$$Y_{ij} = Y'_{ij} + k\epsilon_{ij} \tag{4.2}$$

$$\pi_j \sim \mathcal{N}(\mu_r, \sigma_r^2) \tag{4.3}$$

$$k = \sqrt{\frac{\text{var}(Y'_{ij})}{\text{snr} \times \text{var}(\epsilon_{ij})}} \tag{4.4}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, 1) \tag{4.5}$$

$$\boldsymbol{\beta}^T = [\beta_1, \beta_2, \ldots, \beta_p] \tag{4.6}$$

$$= [\boldsymbol{\beta}_w^T; \boldsymbol{\beta}_{p-w}^T] \tag{4.7}$$

$$X_{ij} = \begin{pmatrix} X'_1 \\ X'_2 \\ . \\ . \\ X'_p \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \mu_x \\ \mu_x \\ . \\ . \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_x & . & \rho_x \\ \rho_x & \sigma_x^2 & . & \rho_x \\ . & . & . & . \\ . & . & . & . \\ \rho_x & \rho_x & . & \sigma_x^2 \end{pmatrix} \right) + s\pi_j J_{n_j,p} \tag{4.8}$$

The parameters of note are s, which is used to scale the random intercepts into the mean of the distributions, and the snr, which affects how much Gaussian noise is being added to the outcome.

In addition to random slopes by cluster, we also consider clustering in the generation of the predictors. We do so by scaling the random intercept by some chosen factor s and adding that to the generated observation (see equation 4.8). Concretely, observations that are in clusters with higher random intercepts are higher on average. This method is equivalent to feeding in a vector of different means as the input to the multivariate distribution, though that is more unwieldy

with the structure of the R command. Thus, some observations have different means in the data generation process.

This adjustment makes the effect of clustering more drastic. Furthermore, in a real-life context, this feature of the simulation approximates a situation where the covariate generation process is inherently different for observations in different clusters. For brevity, we will refer to this parameter of the simulation as the "scale" moving forward.

The Signal-to-Noise (SNR) ratio serves as a proxy for how informative our covariates are. True to the name, a higher SNR corresponds with a smaller amount of Gaussian noise. Thus, for the Gaussian noise present in the data, there is a higher amount of signal between the covariates and the outcome.

We also model measurement noise (separate from the random noise) in our simulation. Although the model is built from the "real" X_{ij}, the observations that are actually included in the resulting data frame have extra noise added to them. This additional noise reflects the measurement error that is present in the original time and labor-intensive data collection process of COP-TOP, which involves categorizing behaviors along somewhat arbitrary lines. We denote the observation that appears in the data that is analyzed as X_{ij}^O. These observations are generated using the following:

$$X_{ij}^O = X_{ij} + \mathcal{N}_p\left(\mathbf{0}_p, \sigma_m^2 I_p\right), \tag{4.9}$$

Using the values in Tables 4.2.1, 4.2.3, and 4.2.2, we run a simulation of 1,000 iterations of each setting of parameters.

Parameter	**Description**	**Value(s) in Simulation**
p	Number of predictors	185
Continued on next page		

Table 4.2.1 – continued from previous page

Parameter	Description	Value(s) in Simulation
w	Number of non-zero slopes	81
d	Default value of non-zero slope	1
μ_x	Mean of covariates (does not take clustering into account)	0
σ_x^2	Variance of randomly generated X_{ij}	1
$\text{cov}(X)$	Correlation between different predictors	0.06
J	Total number of clusters	168
α	Proxy for variation in cluster size	$1/3$
$\bar{n}$	Average number of observations in clusters	3
μ_r	Mean of random intercepts	0
μ_m	Mean of measurement error	0
σ_m^2	Variance of measurement error	1

Table 4.2.1: The description of the parameters and hyperparameters used in the simulation.

The parameters that are controlled by the user and are also held constant for our purposes are detailed in Table 4.2.1. These parameters are determined from Table 3.4.1. Though none of the parameter values in that table are the same for each outcome, they are roughly the same in terms of the type of data they characterize. For example, all of the outcome-specific data sets have a low average covariance between predictors, average cluster sizes between 2 to 3, and a large proportion of estimated coefficients near zero. Furthermore, the number of clusters and predictors are both well over 100, with the number of clusters around or larger than the number of predictors. In the interest of computational intensity, because these parameter values roughly convey the same idea, we often chose a value for the simulation that is easiest to work with. For example, we

choose $J = 168$ for the simulation (instead of 232 or 352 for sake of a faster simulation). For other parameters that presented less clear of a choice, we chose parameters that contributed to conditions similar to the ELS@H data. Notably, for at least some of the combinations of values relating to the clusters, we are able to generate data that has, on average, a roughly equal intra-class correlation to the ELS@H data.

Parameter	**Description**	**Value(s) in Simulation**
σ_r^2	Variance of random intercepts	$[0.00, 0.25, 1.00, 2.25, 4,$ $6.25, 9, 12.25, 16.00]$
s	"Scale" of how much clustering affects the predictors' means	$[0, 1, 2, 3, 4, 5, 6]$
snr	Signal-to-noise ratio	$[0.14, 0.74, 1.34, 1.94, 2.54$ $3.14, 3.74]$

Table 4.2.2: The descriptions and values of the parameters that are varied in the simulation.

Table 4.2.2 lists the parameters that were varied in the simulation. We varied the scale s and the cluster variance σ_r^2 because they tie the clustering to the covariate generation and outcome. Furthermore, because we were also interested in the informativeness of the covariates, we also varied the Signal-to-Noise Ratio (snr).

Parameter	**Description**	**Value(s) in Simulation**
$\boldsymbol{\beta}^T$	Fixed slopes; decomposed into w strong ones and $p - w + 1$ weak ones	$[\boldsymbol{\beta}_w^T; \boldsymbol{\beta}_{p-w+1}^T]$
$\boldsymbol{\beta}_w^T$	Vector of "strong" coefficients	w random draws from a discrete uniform unif(-10, 10)
Continued on next page		

Table 4.2.3 – continued from previous page

Parameter	Description	Value(s) in Simulation
$\boldsymbol{\beta}^T_{p-w+1}$	Vector of "weak" coefficients	$\boldsymbol{\beta}^T_{p-w+1} = (0.5d)^i$, $i = 1, 2, \ldots, p-w+1$
n_j	Number of observations in cluster j	$n_j \sim \text{Unif}(\bar{n}(1-\alpha), \bar{n}(1+\alpha))$
N	Total number of observations in a simulation	$N = \sum_{j=1}^{J} n_j$

Table 4.2.3: The descriptions and values of the parameters that are computed from the parameters in Table 4.2.1.

The computation of the fixed slopes (Table 4.2.3) is inspired by previous simulation studies that compare different variable selection methods (LASSO being one of them) [32, 42]. In fact, we combined several elements of the different settings present in these studies. We chose to have many coefficients close to, but not equal to, zero based on the histograms in Figure 3.4.1, where many of the coefficient values are clustered around zero. Furthermore, there are both negative and positive estimated coefficient values, which is why we drew w times from a Discrete Uniform with a range of -10 and 10. The estimated coefficient values are clearly more extreme than this range. However, with so many predictors, coefficient values that are even slightly large create Y values that are very large in magnitude. This type of data slows down the simulation significantly, as well as creates issues with both types of LASSO.

Figure 4.2.1 is a distribution of a sample generation of the fixed betas as described above. We can see that this distribution does not exactly match the distributions in the data of Figure 3.4.1, but they have in common the clustering of coefficient values around zero, and the occurrence of both positive and negative coefficients. Furthermore,

it is almost certain that the estimated coefficients from the linear regression are not the true coefficient values of the underlying model. Thus, it is acceptable that there is some divergence between the slopes in the simulation and the estimated coefficients from the linear model.

Figure 4.2.1: A sample generation of the fixed betas as specified in Table 4.2.3.

4.3 Performance Metrics

We are primarily interested in the two LASSO regressions' ability to correctly identify which predictors are most important to the model, i.e., which ones have coefficients that have magnitude above $\frac{d}{2}$, where d is the "default" value of the coefficient (see Table 4.2.1). It is important to note that we are not classifying coefficients as zero and non-zero. This classification was originally a consideration, but in a sparse, high-dimensional dataset, it is more likely that the fit model

will estimate coefficients that are quite close to, but not exactly zero. However, variables with a coefficient so close to zero have effectively the same relationship with an outcome variable as a variable with a coefficient of zero. Thus, we classify a coefficient as "negative" (using classification terminology) when it has magnitude less than $\frac{d}{2}$. This dichotomy applies to both the true coefficients of the underlying data and the estimated coefficients. This modified classification is also important to how we generate our data, because $p - w + 1$ of the true coefficients are very close to, but not exactly, zero (due to the nature of the exponential decay).

Furthermore, there are true coefficients from the underlying data that do not have an absolute value to pass the $\frac{d}{2}$ threshold but are not small enough to be effectively zero. These coefficients, which have a magnitude greater than a set threshold of $\frac{d}{8}$ but smaller than $\frac{d}{2}$, are classified as neither true nor false. Thus, we exclude these coefficients from the computation of the F-score.

To that end, sensitivity/recall, specificity, precision, and F1 score are reported. The F1 score, precision, sensitivity/recall, and specificity can be calculated using the following equations:

$$\text{F1} = \frac{2}{\frac{1}{R} + \frac{1}{P}} \tag{4.10}$$

$$P = \frac{TP}{TP + FP} \tag{4.11}$$

$$R = \text{Sensitivity } = \frac{TP}{TP + FN} \tag{4.12}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.13}$$

where TP is the number of true positives, FP is the number of false positives, P is the precision, and R is the recall (after excluding coefficients that have magnitude between $\frac{d}{8}$ and $\frac{d}{2}$.

On a related note, we also reported the root mean squared error

between the predicted coefficients ($\hat{\beta}_j$) and the true coefficients.

Although less important to this thesis, it is also generally useful to compare the predictive abilities of the two LASSO regressions. Thus, the root mean squared error between the predicted outcome and the true outcome was also reported.

Finally, we included the "agreement" between the two models, which is the proportion of predictors that the mixed and regular LASSO yield the same result. In other words, it is the number of predictors that both models deem predictive, plus the number of predictors deemed not predictive by both, divided by the total number of predictors. Again, a covariate is "predictive" of the outcome if the LASSO model assigns it a coefficient with a magnitude larger than $\frac{d}{2}$.

We note that some of these metrics are not particularly informative or appropriate for certain simulated data. In particular, there are edge cases where all of the estimated coefficients have magnitude either all greater than or all less than $\frac{d}{2}$. When we are generating data that is similar to the ELS@H dataset (described in further detail in the section below), we see that the model sometimes estimates coefficients that all have magnitudes below $\frac{d}{2}$. In this case, the precision is NA, since there are neither true positives nor false positives.

5

Results and Discussion

5.1 Primary Results: Agreement (Simulation and Data)

Because we are primarily interested in the informativeness of the covariates, we show results for performance metrics versus the signal to noise ratio. The results in Figure 5.1.1 are under the parameters that roughly approximate the ones in the data (Table 3.4.1).

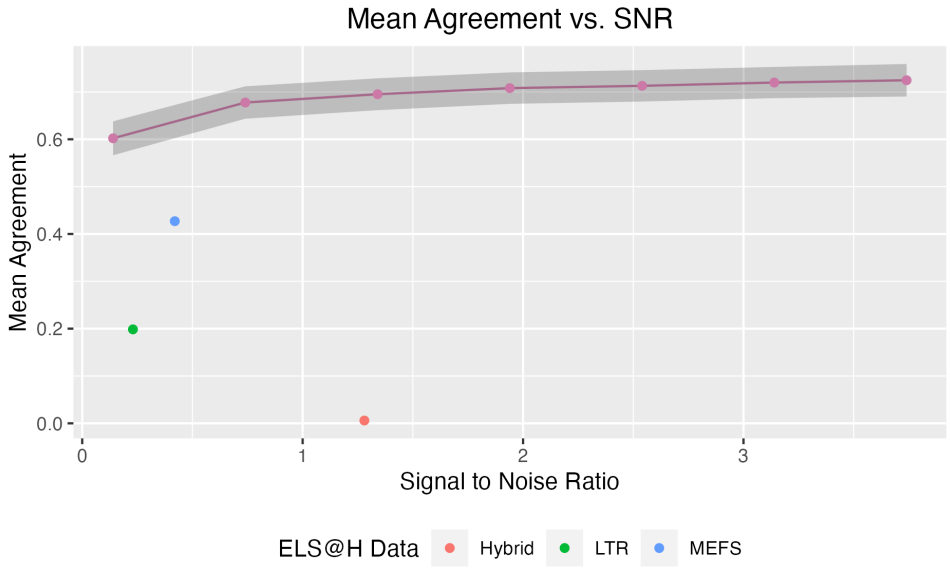

Figure 5.1.1: Mean agreement in classification of coefficients between the fixed and mixed-effects LASSO, versus the signal-to-noise ratio. The agreement values of the data under certain outcomes are also included.

We see that, as the signal-to-noise ratio increases, the mixed-effects and fixed-effects LASSO agree more often on how predictive the covariates are. This agreement is moderately high, with a maximum value of 0.70. We simulated data similar to the ELS@H data, yet there is a stark contrast between the agreement values for the three ELS@H outcomes, and the simulation agreement. The agreement for the Minnesota Executive Function Scale (MEFS) Z-score is the highest of the three, and the PCA-derived hybrid outcome is notably the lowest. The agreement values for all three outcomes are noticeably lower than even the minimum mean agreement value from the simulation.

This contrast between the ELS@H data and the simulated data, as well as the overall purpose of the simulation, motivates us to investigate how this agreement metric varies under other parameters that of the cluster's relationship with the outcome (σ_r^2 and scale, or s).

Specifically, we are interested in how the cluster variance and scale

introduce false relationships between the covariate and outcome. To that end, we also present results with the same parameters that generated the results in Figure 5.1.1, but with different settings of scale and σ_r^2. We have four total settings: scale = 0.75 and $\sigma_r^2 = 0.04$ (the setting in Figure 5.1.1), scale = 0.75 and $\sigma_r^2 = 16$, scale = 6 and $\sigma_r^2 = 0.04$, and scale = 6 and $\sigma_r^2 = 16$. Thus, this data exhibits greater variance in random intercepts, higher correlation between the random intercept values and the covariate means, or both.

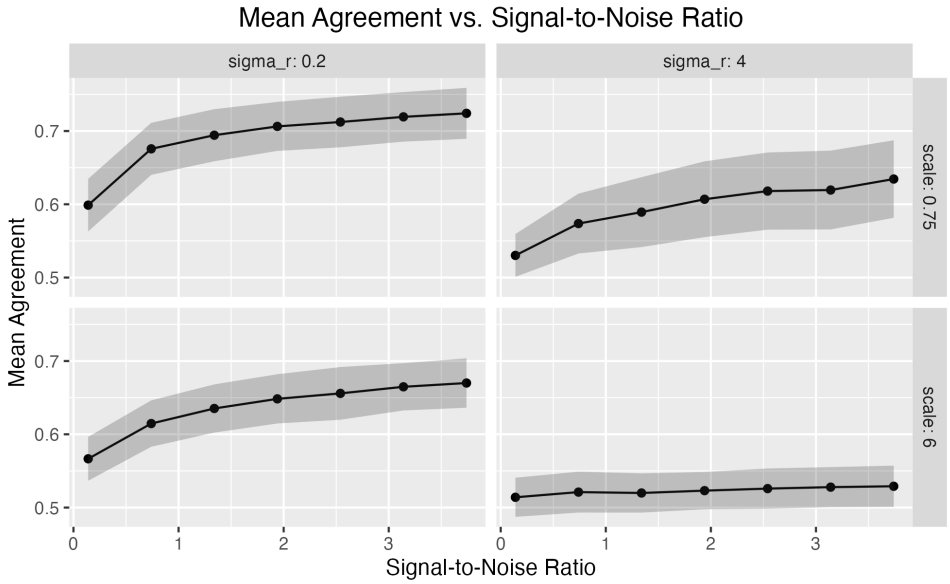

Figure 5.1.2: Mean agreement between fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, under different settings of scale and cluster variance.

As expected, Figure 5.1.2 shows that the agreement between the two flavors of LASSO is lower when the clusters vary more noticeably, in both settings of the scale. Furthermore, holding σ_r^2 constant, the agreement also decreases when the clustering more clearly impacts the covariate generation process. In settings of higher cluster variance and scale (bottom right panel of Figure 5.1.2), the Signal-to-Noise Ratio

does not correspond to an increase in agreement, unlike the other three panels.

Combining these results with those in Figure 5.1.1, it is surprising that the agreement in the ELS@H data is so low when the intra-class correlation for the three outcomes more closely aligns to the simulated data with higher agreement. Referring to Table 3.4.1, the ICCs for the MEFs, Leiter Test, and hybrid outcome are 0.01, 0.13, and 0, respectively. The R^2 values for that same ordering of outcomes are 0.29, 0.18, and 0.58.

SNR	$\mathbf{s = 0.75}$, $\boldsymbol{\sigma_r^2 = 0.04}$	$\mathbf{s = 0.75}$, $\boldsymbol{\sigma_r^2 = 16}$	$\mathbf{s = 6.00}$, $\boldsymbol{\sigma_r^2 = 0.04}$	$\mathbf{s = 6.00}$, $\boldsymbol{\sigma_r^2 = 16}$
0.14	0.04	0.08	0.10	0.12
0.74	0.10	0.28	0.34	0.43
1.34	0.13	0.39	0.47	0.57
1.94	0.15	0.45	0.54	0.66
2.54	0.17	0.49	0.56	0.71
3.14	0.17	0.52	0.64	0.76
3.74	0.18	0.53	0.65	0.79

Table 5.1.1: Average Intra-Class Correlation (ICC) across different Signal-to-Noise ratio values, in different settings of scale and cluster variance.

SNR	$\mathbf{s = 0.75}$, $\boldsymbol{\sigma_r^2 = 0.04}$	$\mathbf{s = 0.75}$, $\boldsymbol{\sigma_r^2 = 16}$	$\mathbf{s = 6.00}$, $\boldsymbol{\sigma_r^2 = 0.04}$	$\mathbf{s = 6.00}$, $\boldsymbol{\sigma_r^2 = 16}$
0.14	0.11	0.10	0.11	0.12
0.74	0.36	0.35	0.39	0.42
1.34	0.48	0.47	0.53	0.57
1.94	0.55	0.55	0.61	0.66
2.54	0.61	0.60	0.64	0.71
3.14	0.64	0.64	0.70	0.76
3.74	0.67	0.65	0.73	0.79

Table 5.1.2: Average R^2 values across different Signal-to-Noise ratio values, in different settings of scale and cluster variance.

The results of the simulation suggest that the fixed and mixed-effects LASSO will diverge in settings related to a high ICC; however, the data does not fall under that category. Thus, the low agreement between the two may not be related to the cluster variance and correlation with the covariance generation. Instead, perhaps the inconsistency across different variants of LASSO are due to the true weakness of the predictors, or another underlying characteristic of the data that the simulation does not model well. Specifically, there may be other ways besides the cluster variance and the scale to adjust the simulation to achieve the low ICC that this thesis does not account for; given more time, this direction would be worth exploring.

5.2 Primary Results: Classification (Simulation Only)

To better understand other trends in these different settings of σ_r^2 and s, we also examine the F-score, an indicator of how well each LASSO type correctly classifies variables as predictive. In Figure 5.2.1, we see that the fixed-effects LASSO exhibits a lower F-score in both setting of cluster variation. However, in all settings but the one with scale = 6 and $\sigma_r^2 = 16$, the fixed-effects LASSO is able to approach the F1 score of the mixed-effects LASSO with an increase in SNR. It is also worth noting that the F-scores for both LASSO models are not particularly high (even within one standard deviation, they have approximate scores of at most 0.60).

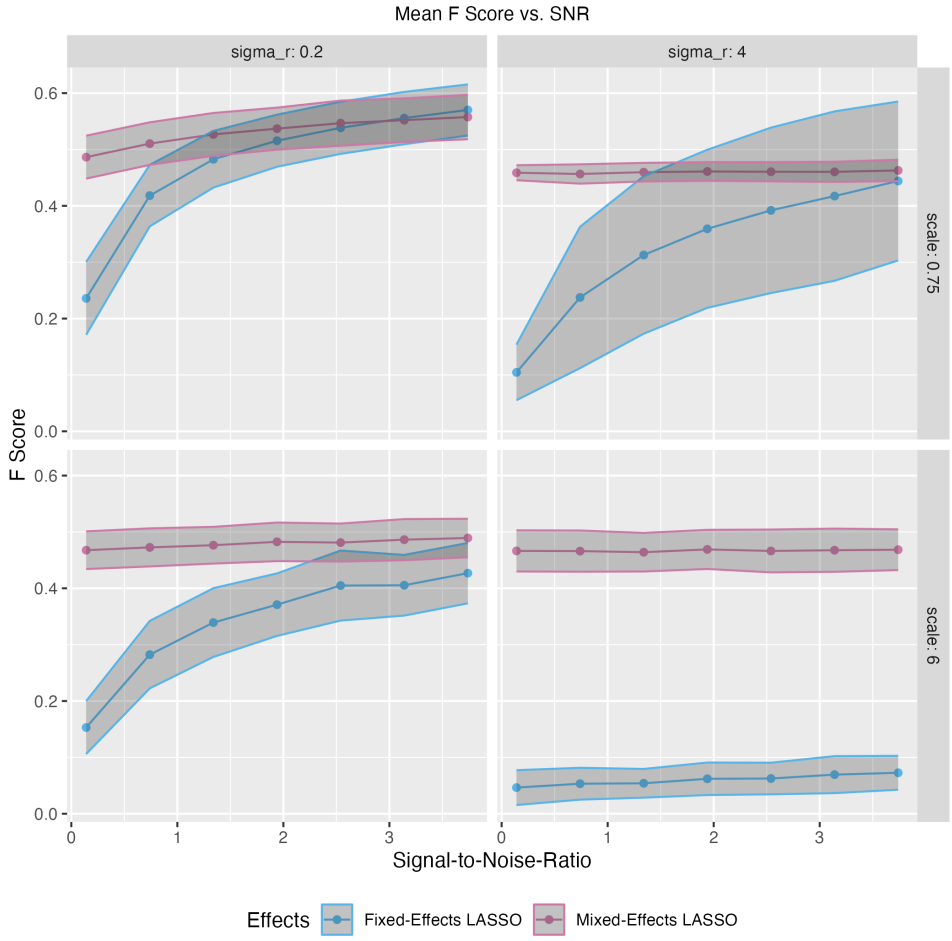

Figure 5.2.1: Mean F-score between fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, under two different settings of cluster variation.

We also present plots of results while varying the variance in cluster and scale. We examine these results in two different settings of SNR, based on the approximate minimum and maximum values in Table 3.4.1: 0.30 (actually 0.23 in the table) and 1.28. We define these bounds to be consistent with simulating data as close to the ELS@H data as possible.

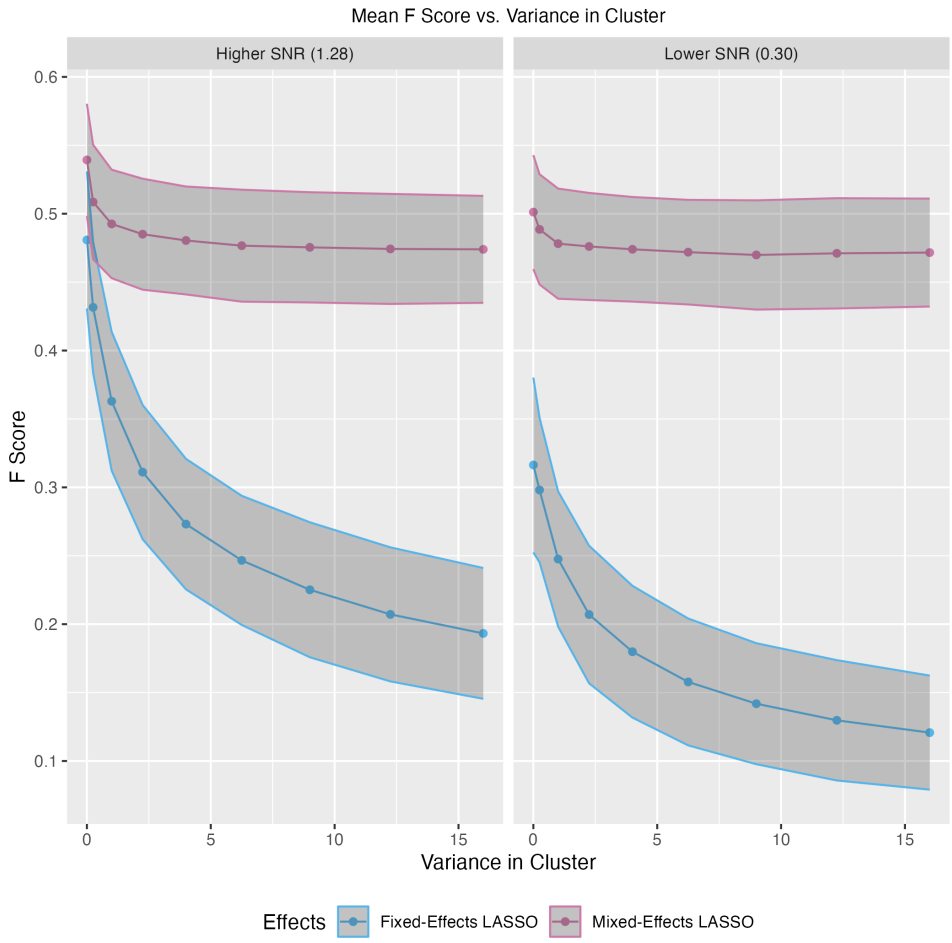

Figure 5.2.2: Mean F-score of the fixed and mixed-effects LASSO, in relation to the variance of the random intercepts (`sigma_r`^{2} of the data).

In Figure 5.2.2, it is evident that from the standpoint of an F-score, the mixed-effects LASSO is more resilient to an increase in cluster variance. However, the overall low F-score is still a point of curiosity, so we also more closely examine this classification through the total number of predictive coefficients, precision, and recall.

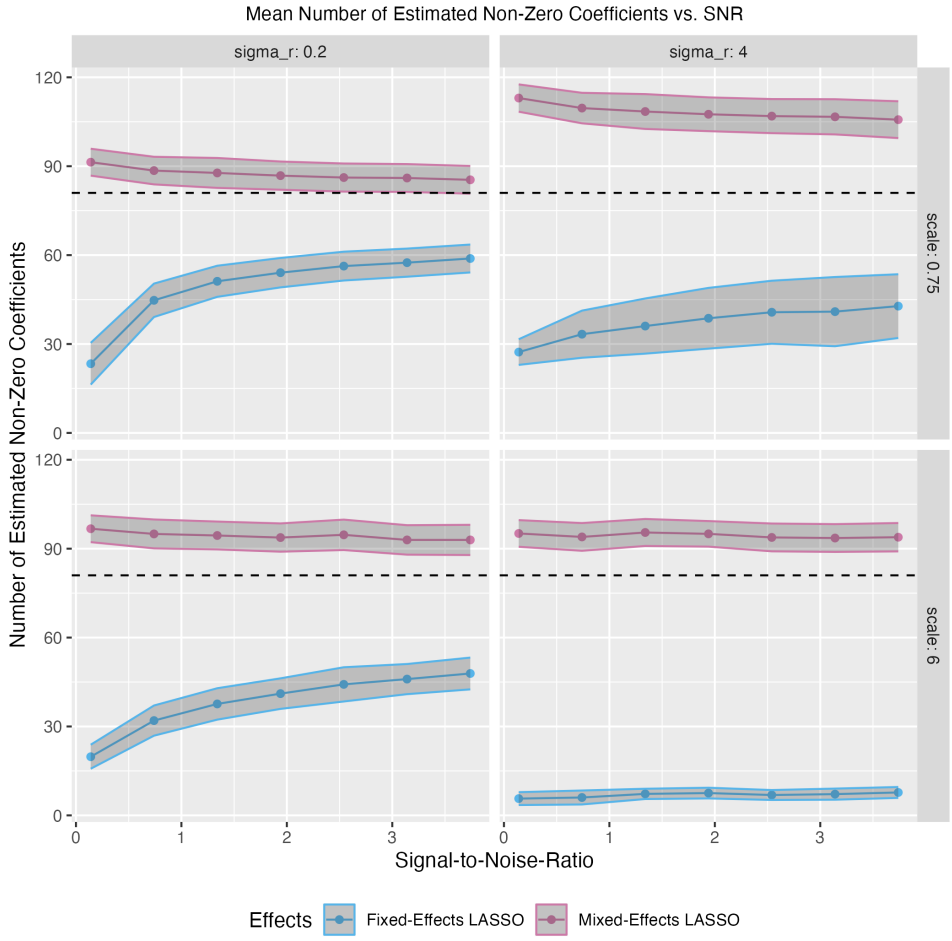

Figure 5.2.3: Mean number of non-zero coefficients of the fixed and mixed-effects LASSO, in relation to the signal-to-noise ratio of the data, over different settings of cluster variance and scale. The true number of non-zero coefficients, $w = 81$, is indicated by the dotted horizontal line.

Figure 5.2.3 shows that as the clusters vary more, the fixed-effects LASSO shrinks more coefficients to zero. This pattern would explain the low F-score for the fixed-effects LASSO in the corresponding setting, since there are simply fewer opportunities for the fixed-effects LASSO to score a true positive. In contrast, the mixed-effects LASSO demonstrates the opposite pattern, classifying relatively the same

amount of coefficients as predictive regardless of the changes in SNR. Though mixed-effects LASSO consistently overestimates the number of predictive covariates, it estimates more closely to the number of true predictive covariates. However, this closeness in predicted versus true number of strong coefficients does not necessarily imply that the mixed-effects LASSO is classifying the corresponding coefficients correctly. We must look at the precision and recall to better understand the low F-score.

Figure 5.2.5 of the precision shows that though the the mixed-effects LASSO classifies an amount of coefficients closer to the true value, the model is not particularly successful in classifying the specific coefficients *correctly*. In other words, the mixed-effects LASSO seems to be tagging many coefficients as predictive, but this is not necessarily accurate. The mixed-effects LASSO does exhibit a higher recall than its fixed-effects counterpart (Figure 5.2.5), and we can connect this with Figure 5.2.3. Because the mixed-effects LASSO classifies more coefficients as predictive, it has more opportunities to be correct.

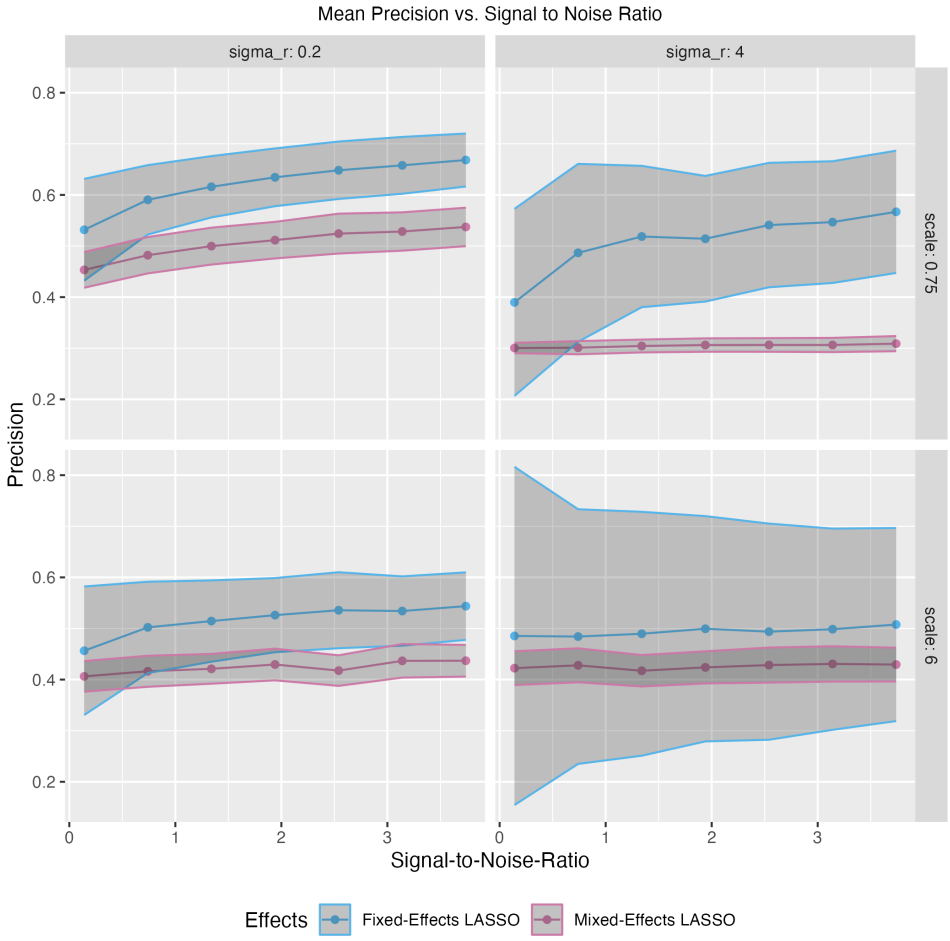

Figure 5.2.4: Mean precision between fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, under two different settings of cluster variation.

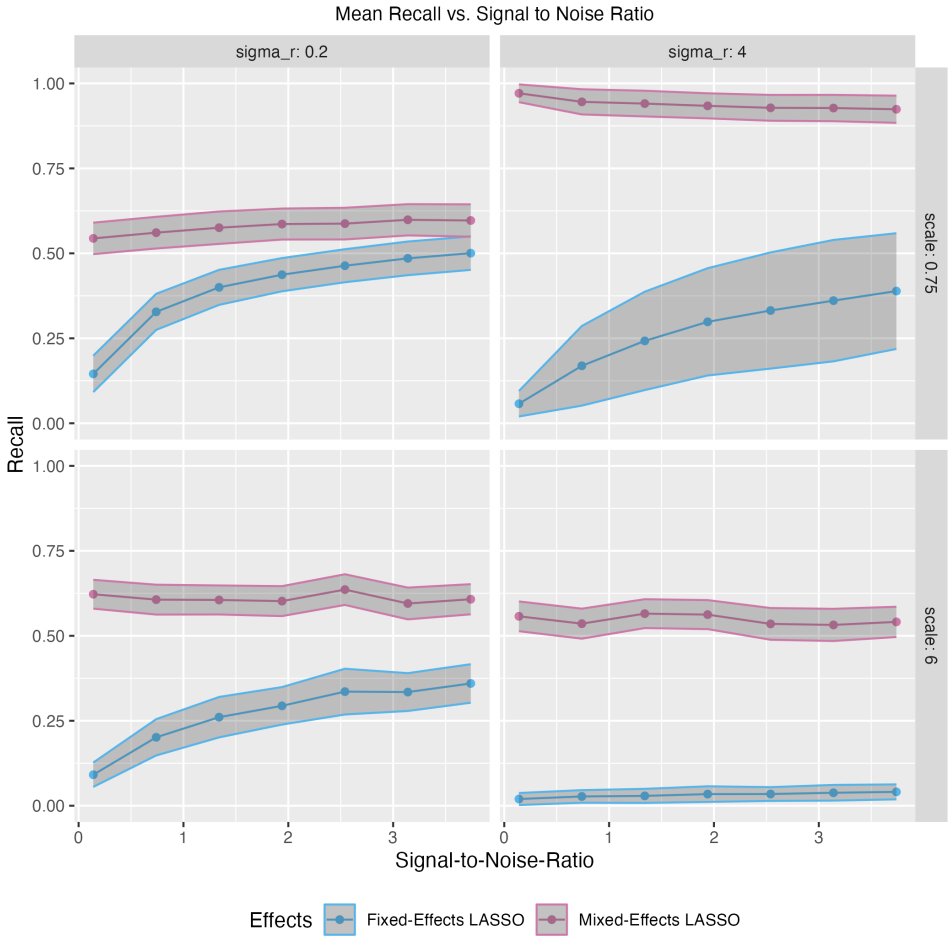

Figure 5.2.5: Mean recall between fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, under two different settings of cluster variation.

From all figures in this section, we see that the outcomes for the mixed-effects LASSO experience fairly small changes in response to a change in parameter (whether it be SNR, cluster variance, or scale). In contrast, the fixed-effects LASSO is more sensitive to changes in parameters.

To summarize this subsection, we see that there are some caveats to the higher F-score of the mixed-effects LASSO. Specifically, its higher F-score (than the fixed-effects LASSO) can mostly be attributed to its

somewhat indiscriminate classification of variables as predictive, especially in more extreme cluster settings (higher scale and σ_r^2). This behavior leads to a higher recall than its fixed-effects counterpart. Relatedly, the mixed-effects LASSO does worse in precision, or it is not particularly accurate in identifying the truly predictive covariates (among the ones it does classify as predictive). Moreover, both types of LASSO exhibit low F-scores. This overall perspective calls into question both models' abilities to capture both the predictive covariates (recall) and be accurate with the covariates it does flag as predictive (precision).

5.3 Secondary Results: Practical Matters

Although not the primary purpose of the simulation, it is also of general interest how the mixed-effects and fixed-effects LASSO compare in predicting outcomes and estimating coefficients. To provide a slightly different perspective, we look at how these metrics vary against the cluster variance (all other plots are located in the Appendix). In addition, it is of practical importance to Early Childcare Providers which covariates were selected by the different types of LASSO.

5.3.1 Accuracy

The below results are only from the simulation (and not a comparison between the data and the simulation), mainly because we do not have the "true" parameter values, such as the actual values of the coefficients. Instead, we used information from a bare-bones linear regression of the already-generated ELS@H data and have little knowledge about the "true" data-generating process. The outcomes explored in this section are the prediction RMSE (accuracy of predicting outcomes) and the estimated coefficients RMSE (accuracy

of estimating the true model). Only plots representing the overall trends were included; please see the Appendix for more results involving other parameters.

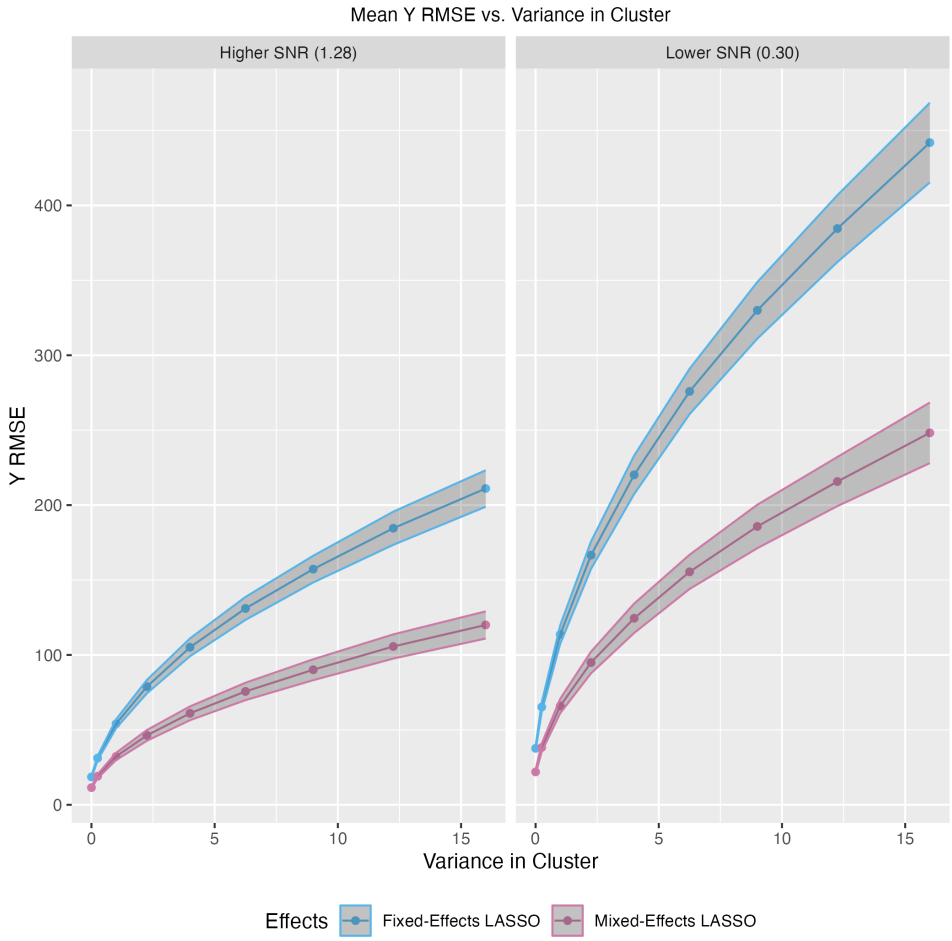

Figure 5.3.1: Mean prediction RMSE of the fixed and mixed-effects LASSO, in relation to the variance of the clusters.

On average, the mixed-effects LASSO consistently does a better job of predicting the outcome across different cluster variance values (Figure 5.3.1). However, we should note that the RMSE from the mixed-effects LASSO is still rather large for larger values of variance.

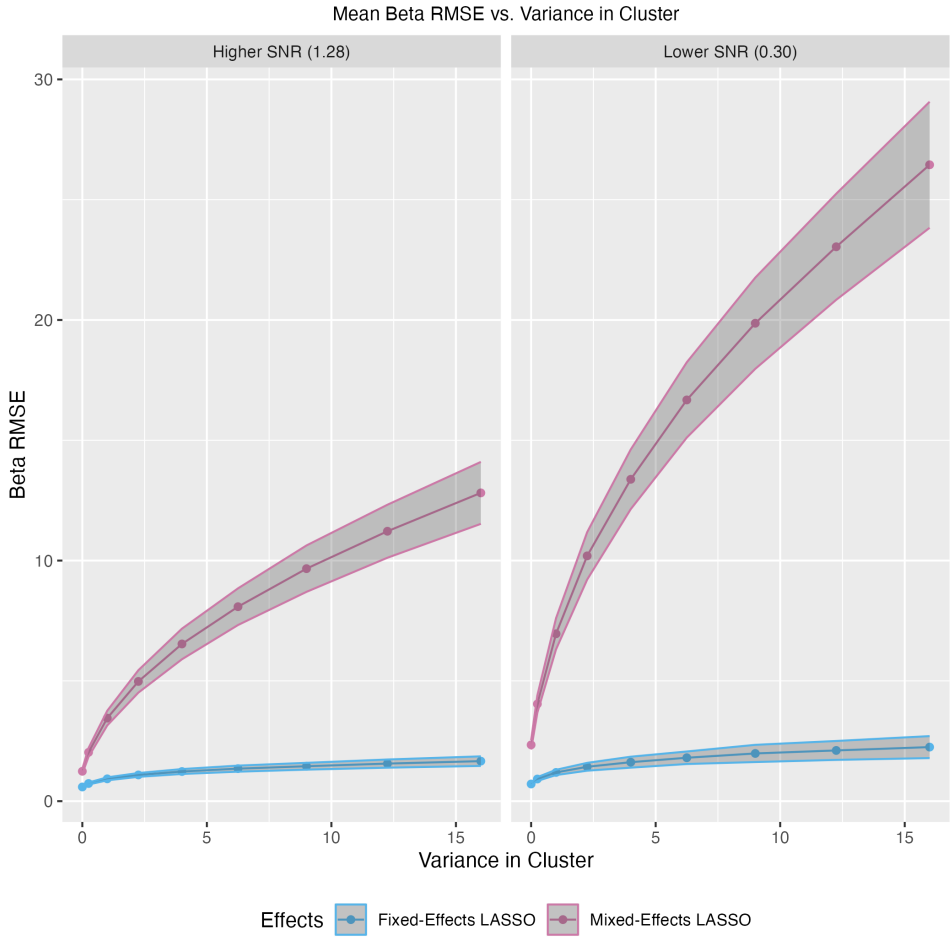

Figure 5.3.2: Mean estimated coefficient RMSE of the fixed and mixed-effects LASSO, in relation to the cluster variance of the data.

To lend more context to these prediction RMSE values, we also examine the RMSE of the estimated coefficients. The mixed-effects LASSO underperforms in estimating the actual values of the coefficients when compared to the fixed-effects model (Figure 5.3.2). The contrast of the two imply that the mixed-effects LASSO would not perform well out-of-sample, especially in the presence of a lower SNR. Its prediction accuracy does not come from estimating the true values of the coefficients. In addition, as expected, both the fixed and

mixed-effects LASSO do a better job of estimating the true coefficients and predicting the outcomes in the presence of a higher SNR.

Both of these trends also hold true when holding the variance in cluster constant and varying the scale (the corresponding plots are included in the Appendix).

These results suggest that the mixed-effects LASSO, in this particular setting of data, may not have the desired result. This conclusion, coupled with the previous suggestion that the weak predictive signal in the ELS@H data is not caused by clustering, calls into question whether the computational intensity of mixed-effects LASSO is worth the trouble for this data set. In particular, as detailed in previous sections, the process for picking a regularization parameter through cross-validation is particularly painstaking for mixed-effects LASSO, and this version of the mixed-effects LASSO is ill-equipped to work with highly correlated covariates. Thus, this simulation study suggests that considering the clustering of the students into providers may not be very important to decoding the lack of predictive signal. Moreover, we also need to consider whether considering the clustering of the students in this data set is advantageous in general, given similar levels of accuracy despite large differences in computational intensity.

5.3.2 Variables Selected

As mentioned in the earlier section, in the simulation, the mixed-effects LASSO often overestimates the true number of non-zero covariates, while the fixed-effects LASSO severely underestimates. This section compares this result to the selection of variables in the actual data. Though we do not actually know the true number of non-zero coefficients because it is impossible to find the true data-generating process, we can at least examine the relative amount

of coefficients that each shrink to zero. We also highlight the predictors that both the mixed and fixed-effects LASSO deemed as predictive. A full list is included in the Appendix.

Outcome	Mixed LASSO	Fixed LASSO
MEFS	141	10
Leiter	130	21
Hybrid	115	0

Table 5.3.1: The number of non-zero coefficients estimated by the mixed and fixed-effects LASSOs, across the different outcomes in the data.

From Figures 5.2.3 and A.1.9, we see that this pattern of the mixed-effects LASSO shrinking fewer coefficients to zero persists across different settings of Signal-to-Noise ratio and cluster variance. Thus, if the ELS@H data's cluster variance is not correctly specified, or the signal-to-noise ratio definition in the simulation is inconsistent with what is happening in the data, these two results are likely consistent.

There were no predictors that were consistently selected by the mixed-effects and fixed-effects LASSO across all three outcomes. With the exception of the hybrid outcome (where the fixed-effects LASSO shrank all coefficients to zero), for each outcome, all of the variables that the fixed-effects LASSO selected were also selected by the mixed-effects LASSO. There was one predictor that was selected by both MEFS's LASSO and Leiter's LASSO was `o_t_tone_o_N`. This predictor represents the proportion of the time (out of the total number of sweeps observed for a particular student) that a teacher or assistant looks "displeased" or displays "annoyance or disappointment" [36].

6 Conclusion

In conclusion, the simulation study suggests that the clustering of the students alone is not to blame for the lack of predictive signal in the data. However, this conclusion is not without its caveats, and we address any limitations and future improvements of this project.

6.1 Limitations

First and foremost, the ELS@H data cannot be perfectly replicated by any simulation. For example, we had to generalize the form of the coefficients as specified in Table 4.2.3, but it is clear from the histograms in Figure 3.4.1 that there may be several covariates with coefficients that lie outside of the range, and even for those that lie in the same range, they may not necessarily have the same distributions.

Another example is that the covariates in the simulation were all generated from a Normal distribution, when real-life data are perhaps best approximated by other distributions. Overall, the data are not generalizable to this type of simulation, and other parameters also had to be decided on a somewhat arbitrary basis to allow the simulation to run.

In addition, the current implementation of mixed-effects LASSO is ill-equipped to work with highly correlated or perfectly collinear predictors. As mentioned in the Data section, this problem required data pre-processing along somewhat arbitrary lines. On a similar note, there is a lack of built-in functions for the mixed-effects LASSO that perform certain computations as efficiently as the fixed-effects LASSO.

6.2 Future Directions

The ELS@H data set is quite rich, and there are many further directions we can explore. For one, there are eight other outcomes in the data, and it may be worth stepping through a similar process for the other outcomes. Furthermore, data for years later than the first two years of the study are available, though there may not be as many observations.

The simulation can also be made more flexible. For example, it is worth considering distributions other than the Normal, or other "types" of coefficients besides the ones explored in previous literature. There are also other levers of the simulation that can be changed. For example, measurement noise was not explored very deeply in the simulation, but members of ELS@H have attested that it does play some role in the measurements of COP-TOP.

The collinearity of the data is unavoidable, and there exist fixed-effects methods that better accommodate this issue. Though

LASSO is attractive for its computational efficiency and ability to completely shrink certain coefficients to zero, it was not necessarily designed to work with highly correlated data. In contrast one of Ridge's original design purposes was to address collinearity. The Elastic Net balances both LASSO's predictor selecting ability and Ridge's infrastructure to work with collinearity, by combining the Ridge and LASSO penalties [43]. Thus, another future direction is to step through a similar study with a multilevel Elastic Net, which is even less developed in the literature than the regular multilevel LASSO.

6.3 Final Thoughts and Recommendations

This project does not seek to discount random effects as a whole, as they are a fundamental part of analyzing education data. Rather, through simulation, we provide a new way to understand the source of a weak predictive signal in an important early childhood data set. Specifically, we have found some evidence that the clustering in the data does not contribute to the weak predictive signal. In addition, we show some shortcomings of considering mixed effects in variable selection. Thus, considering clustering by provider may not be necessary when performing variable selection on the data set. For practical next steps, we encourage future users of this data set to consider these described advantages and disadvantages of clustering by provider from a prediction, classification, and computational standpoint.

A

Appendix

All code used for this project can be found in this GitHub repository. Some of this code was written in collaboaration with Jonathan Seiden and the Early Learning Studies at Harvard, whom I thank profusely for their help!

A.1 Simulation Results

A.1.1 Signal-to-Noise Ratio

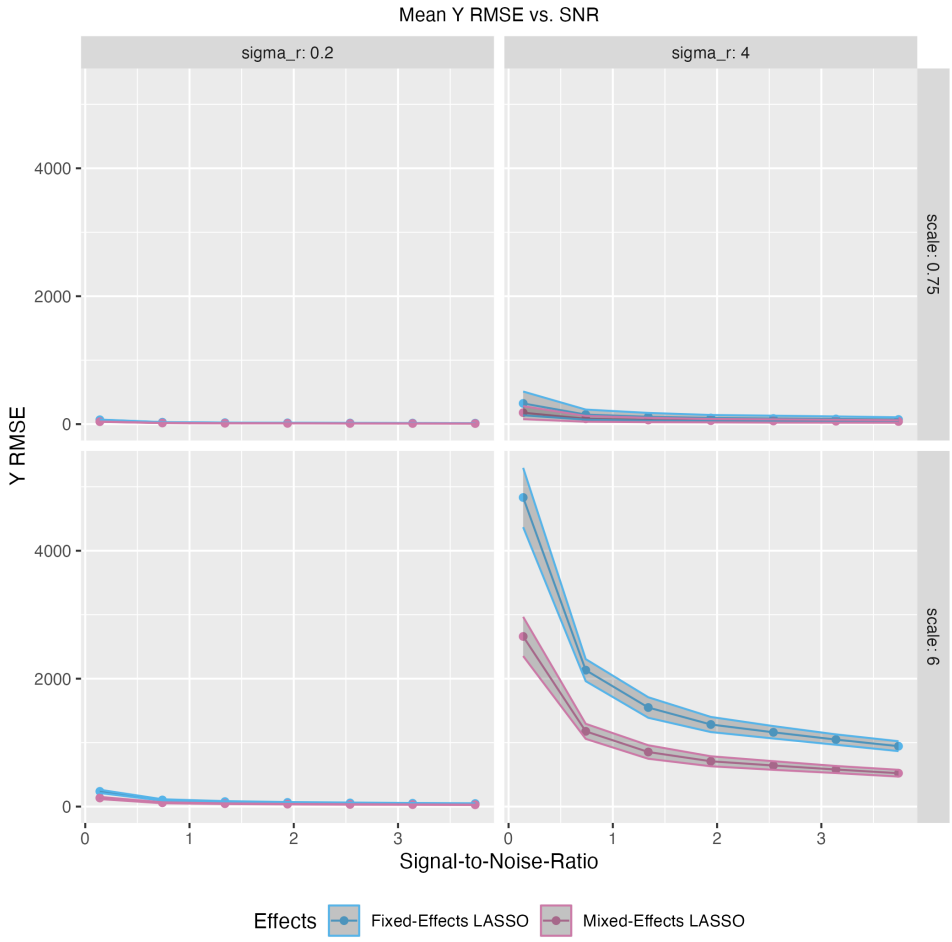

Figure A.1.1: Prediction RMSE of the fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, over different cluster settings.

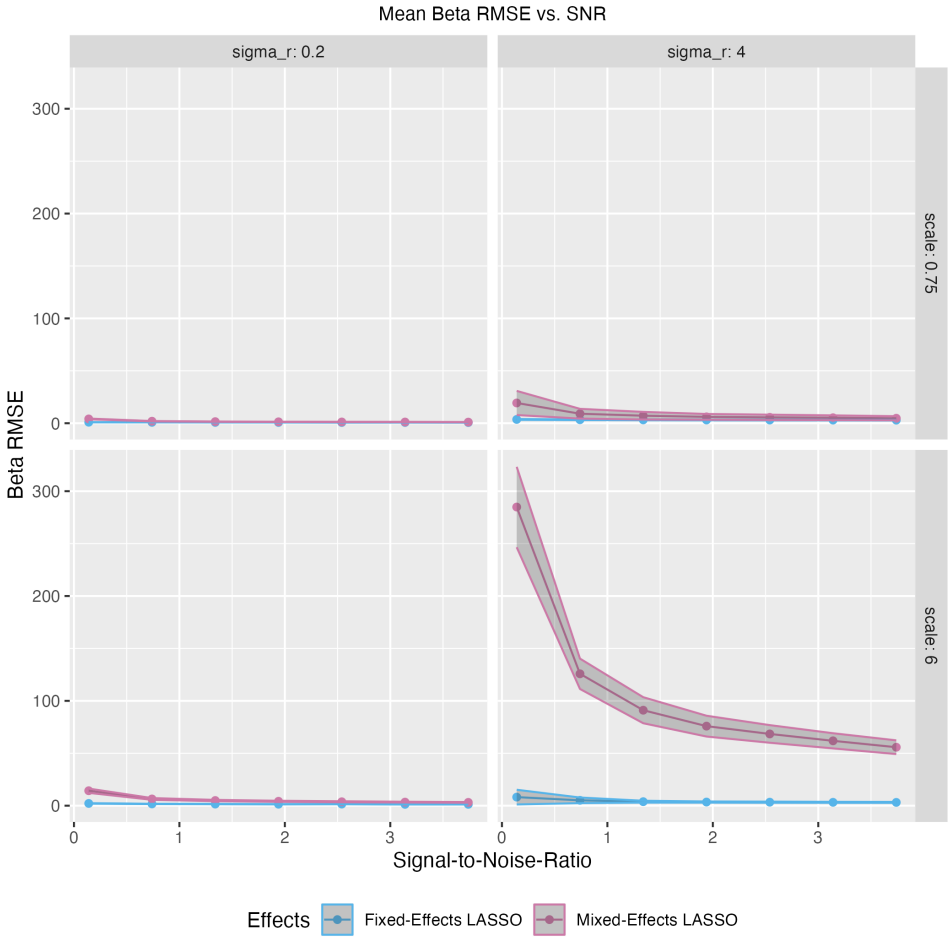

Figure A.1.2: Estimated Coefficient RMSE of the fixed and mixed-effects LASSO, in relation to the Signal-to-Noise Ratio, over different cluster settings.

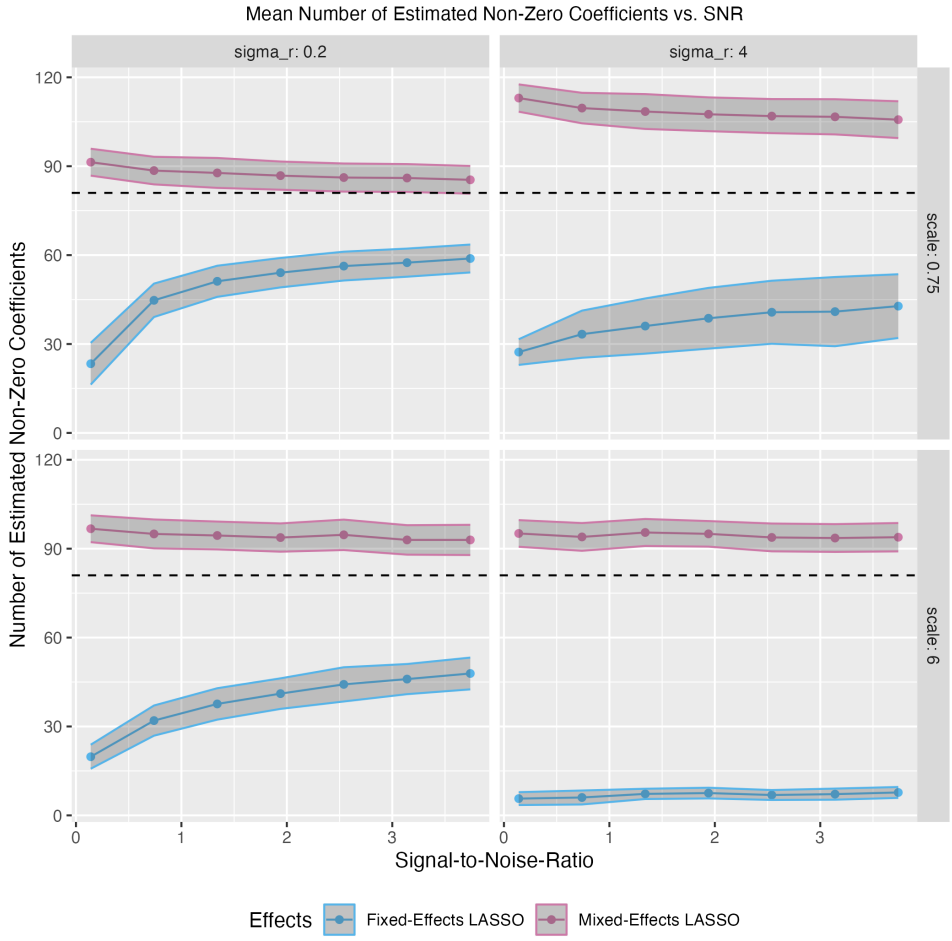

Figure A.1.3: Mean number of estimated non-zero coefficients of the fixed and mixed-effects LASSO, in relation to the Signal-to-Noise ratio, over different cluster settings. The true value of non-zero coefficients is $w = 81$, denoted by the dotted line.

A.1.2 Cluster Variance

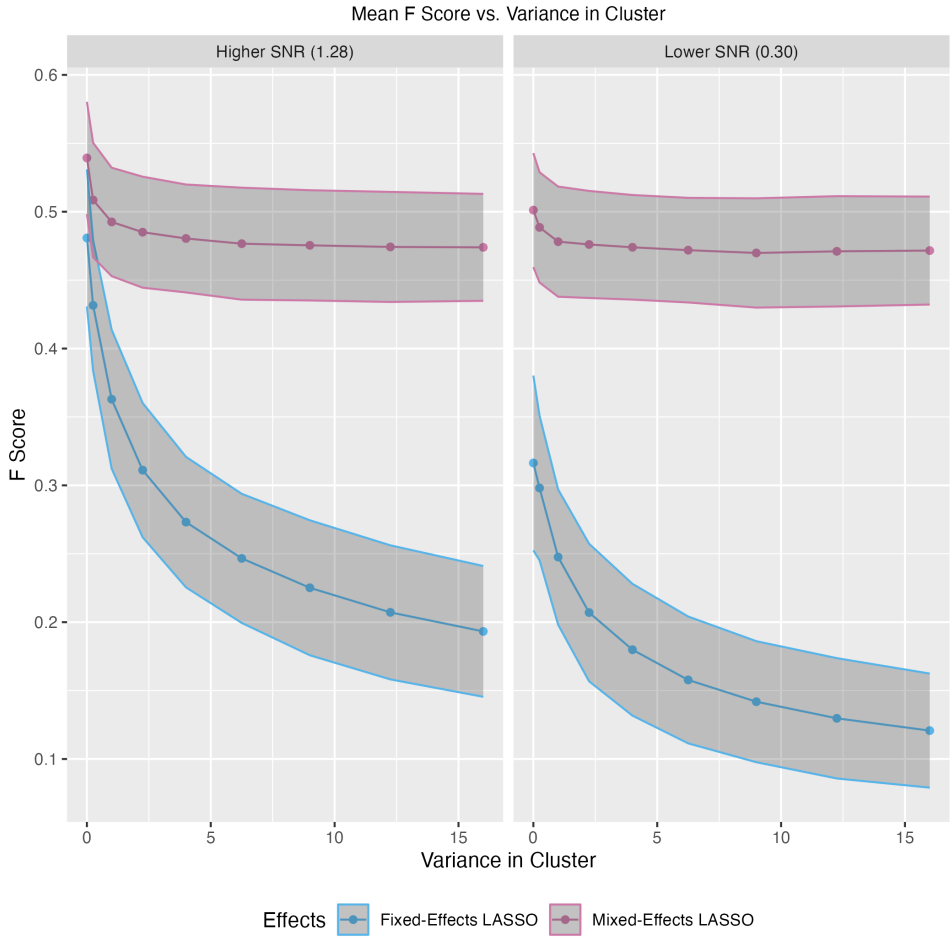

Figure A.1.4: Mean F-score of the fixed and mixed-effects LASSO, in relation to the variance of the random effects.

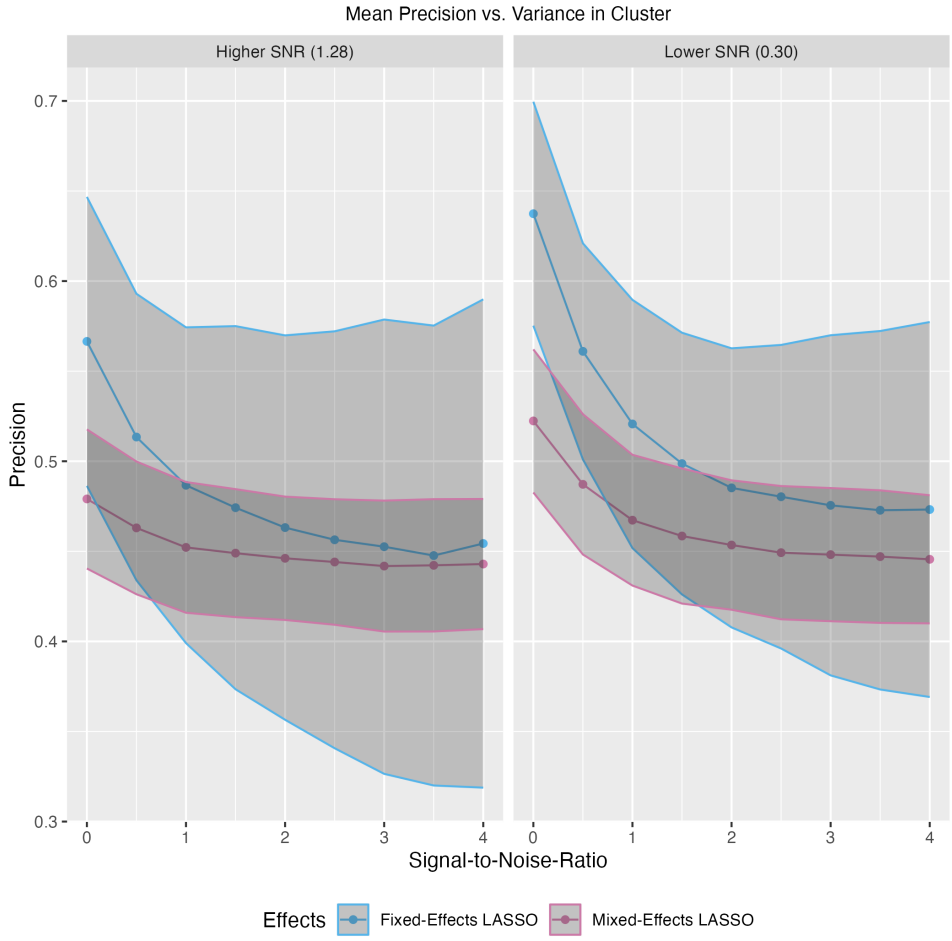

Figure A.1.5: Mean precision of the fixed and mixed-effects LASSO, in relation to the variance of the random effects.

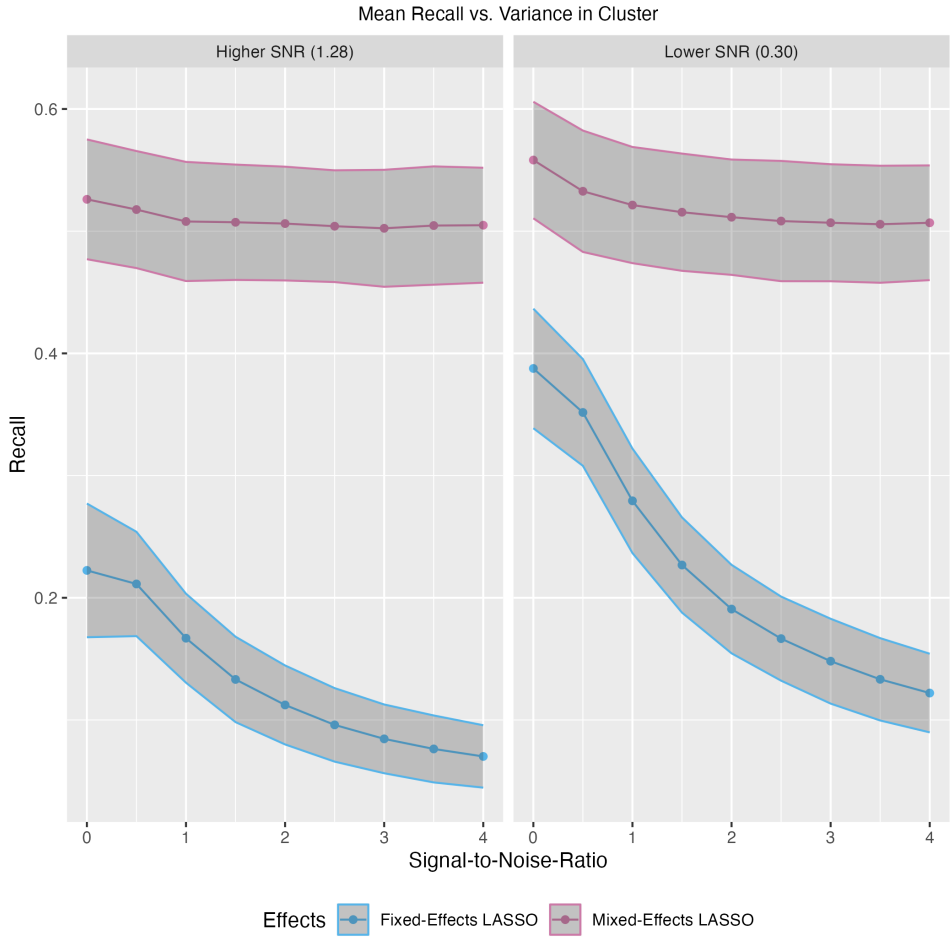

Figure A.1.6: Mean recall of the fixed and mixed-effects LASSO, in relation to the variance of the random effects.

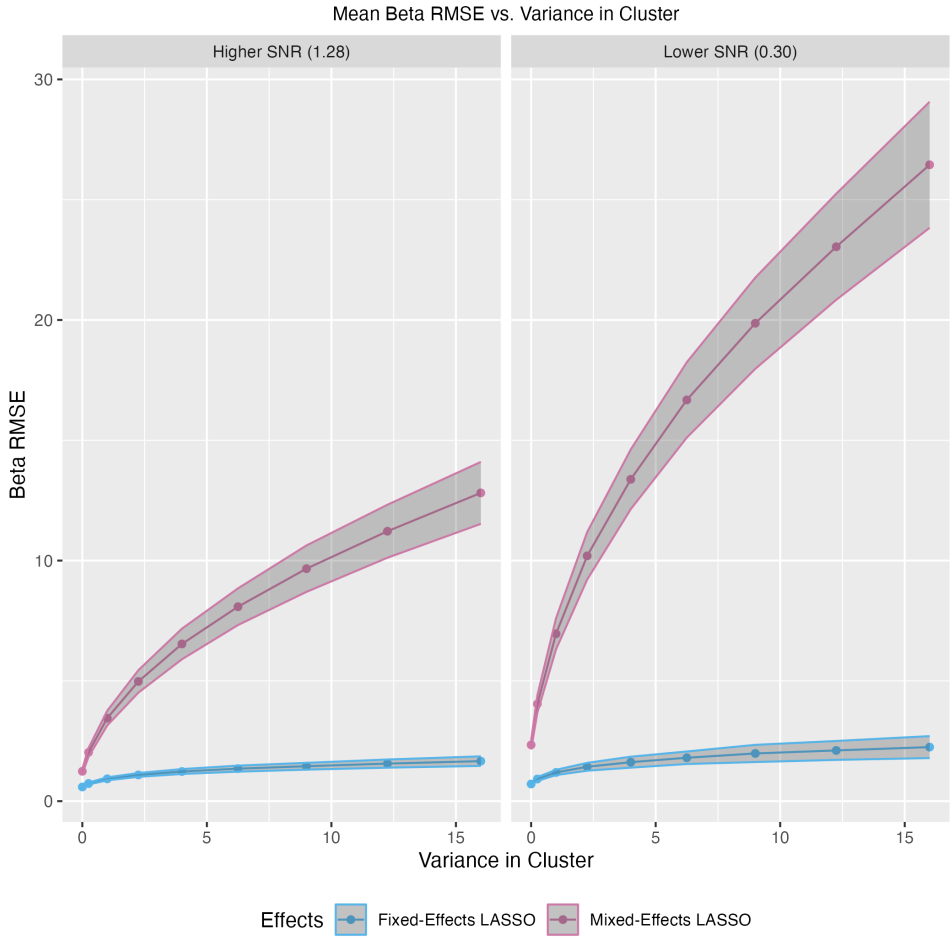

Figure A.1.7: Mean estimated coefficient RMSE of the fixed and mixed-effects LASSO, in relation to the variance of the random effects.

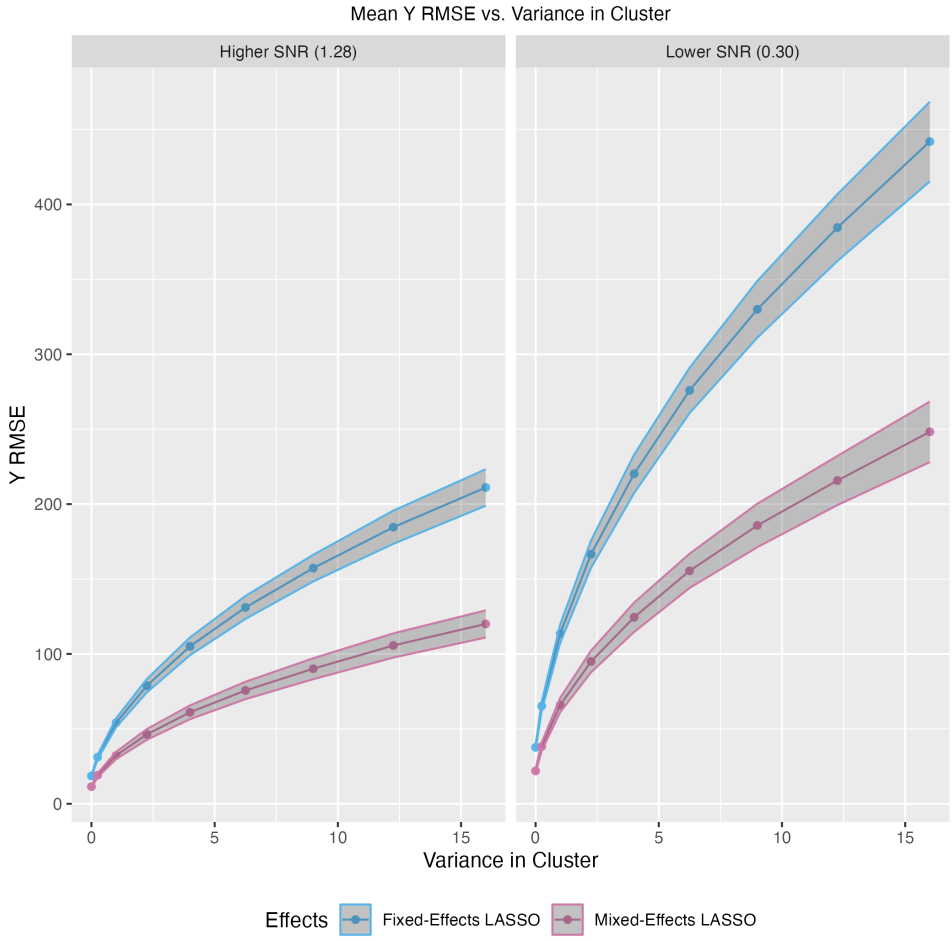

Figure A.1.8: Mean prediction RMSE of the fixed and mixed-effects LASSO, in relation to the variance of the random effects.

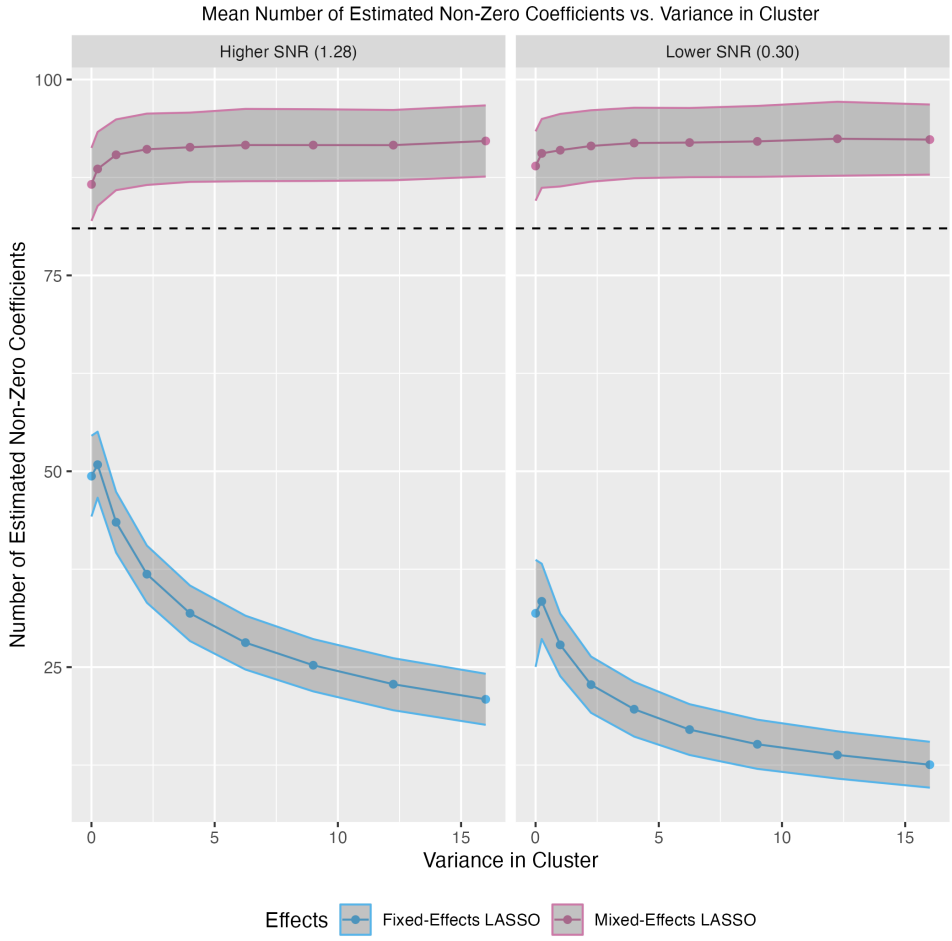

Figure A.1.9: Mean number of estimated non-zero coefficients of the fixed and mixed-effects LASSO, in relation to the variance of the random effects. The true value of non-zero coefficients is $w = 81$, denoted by the dotted line.

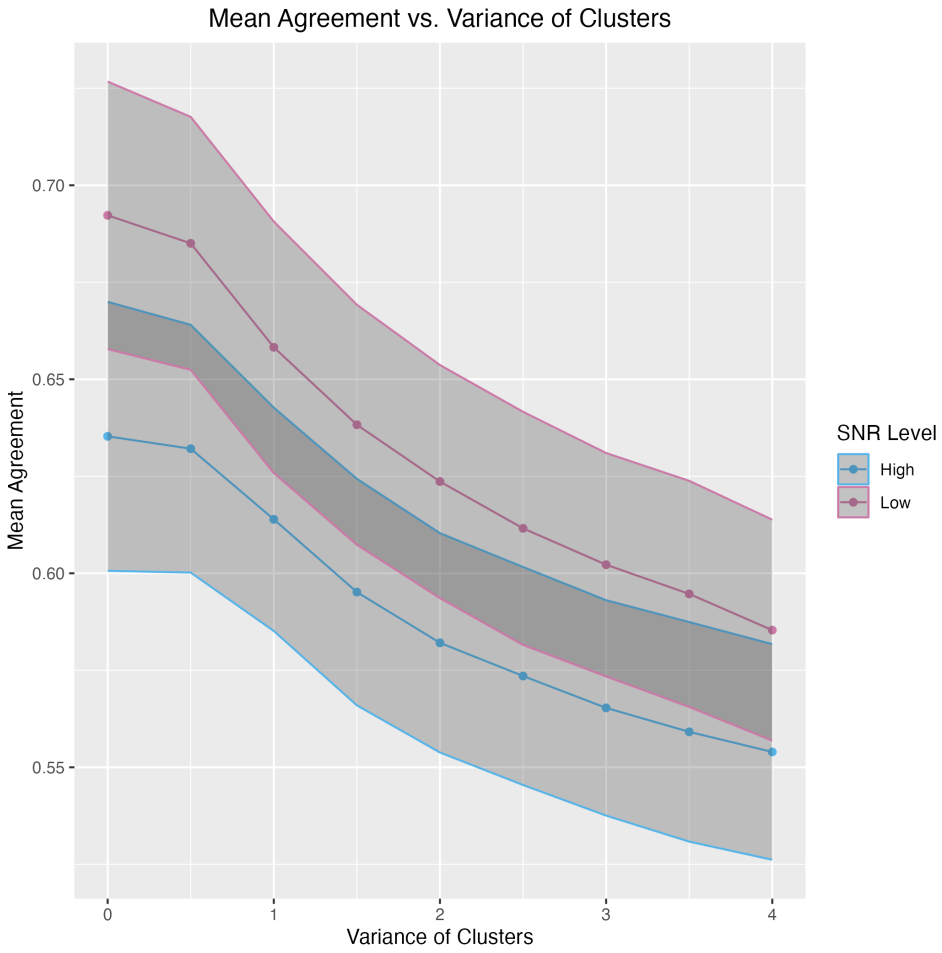

Figure A.1.10: Mean Agreement between fixed and mixed-effects LASSO, in relation to the variance of the clusters.

A.1.3 Scale

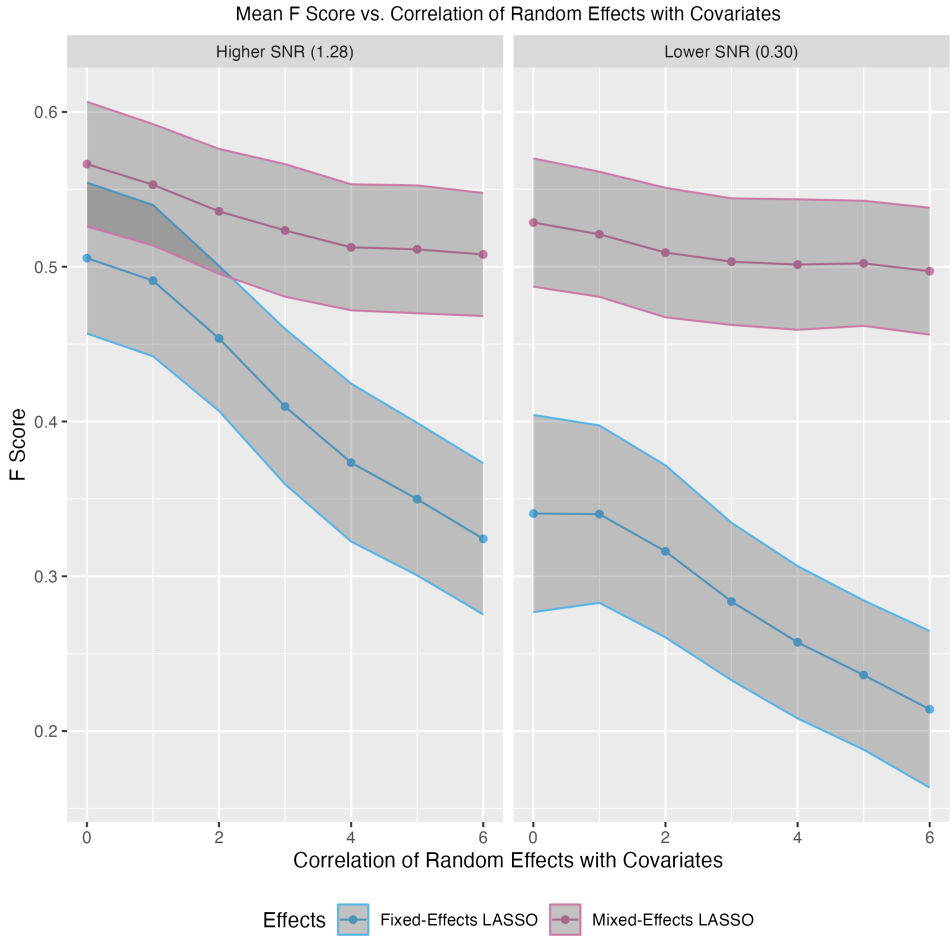

Figure A.1.11: Mean F-score of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation.

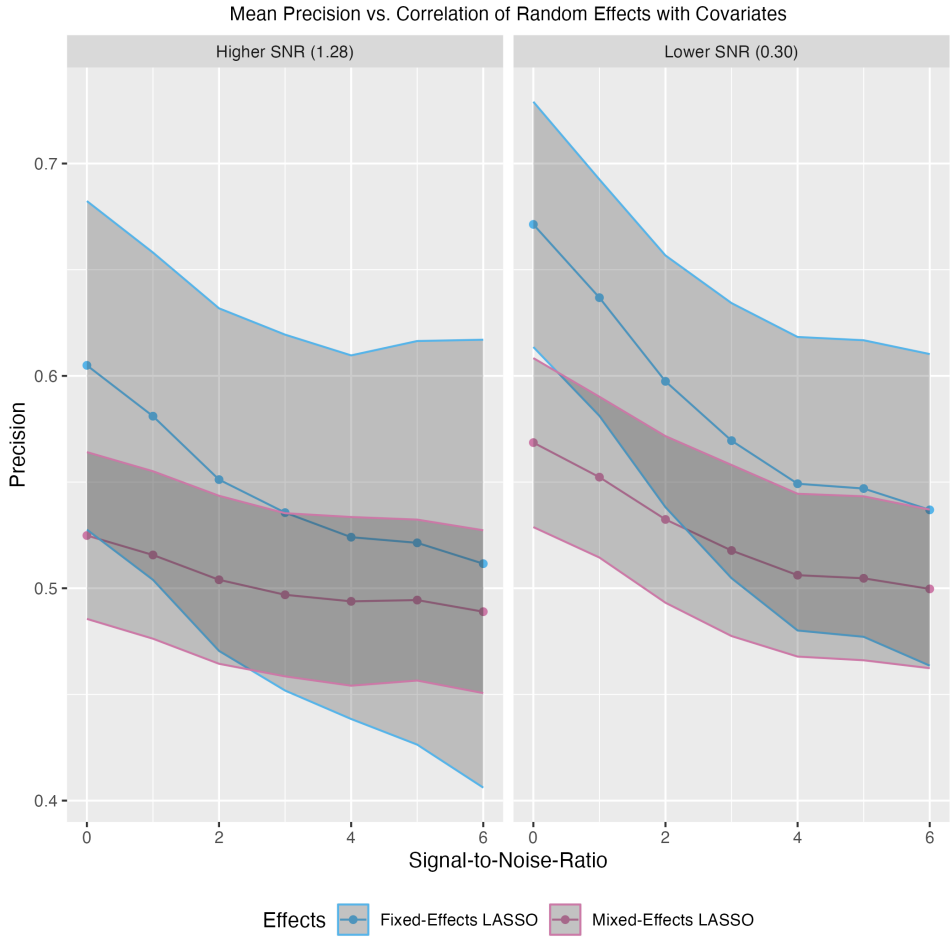

Figure A.1.12: Mean precision of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation.

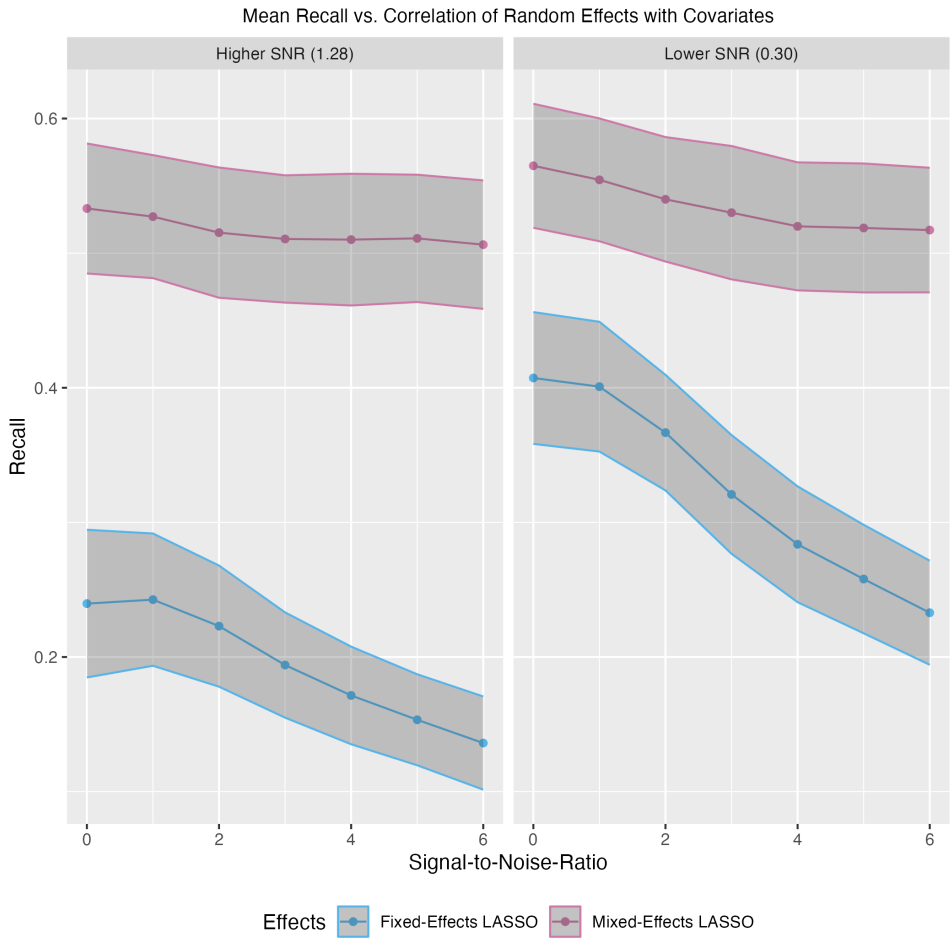

Figure A.1.13: Mean recall of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation

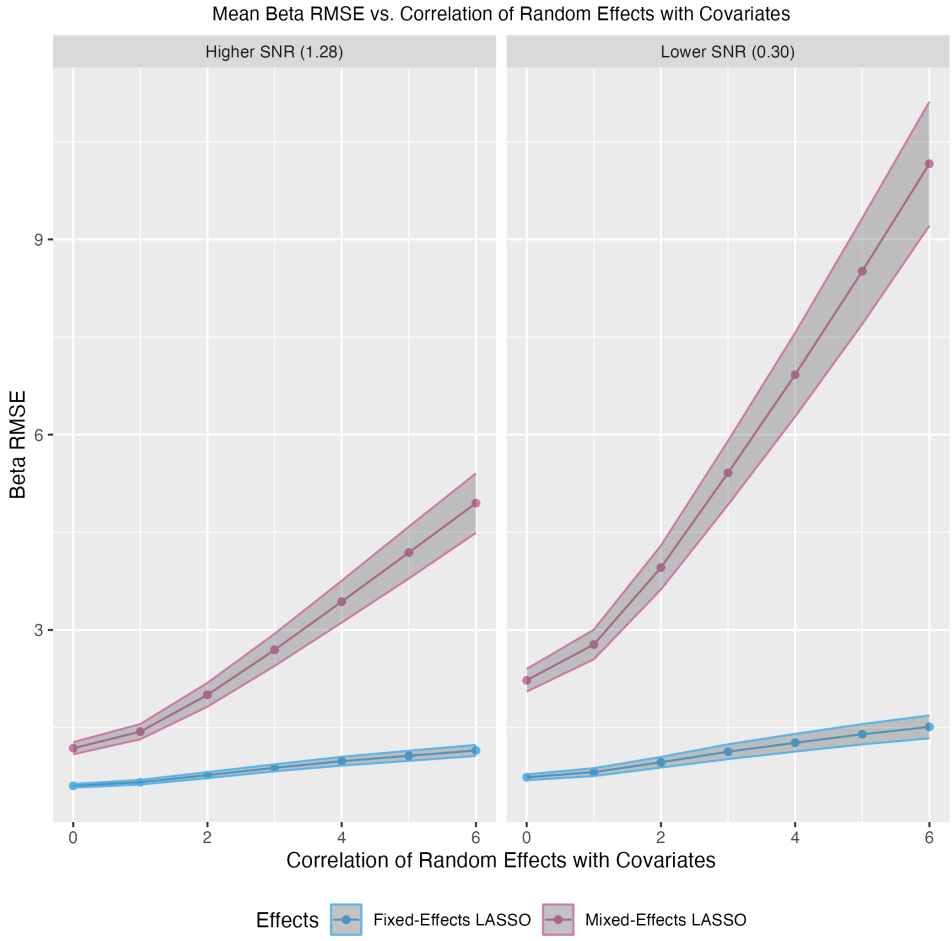

Figure A.1.14: Mean estimated coefficient RMSE of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation.

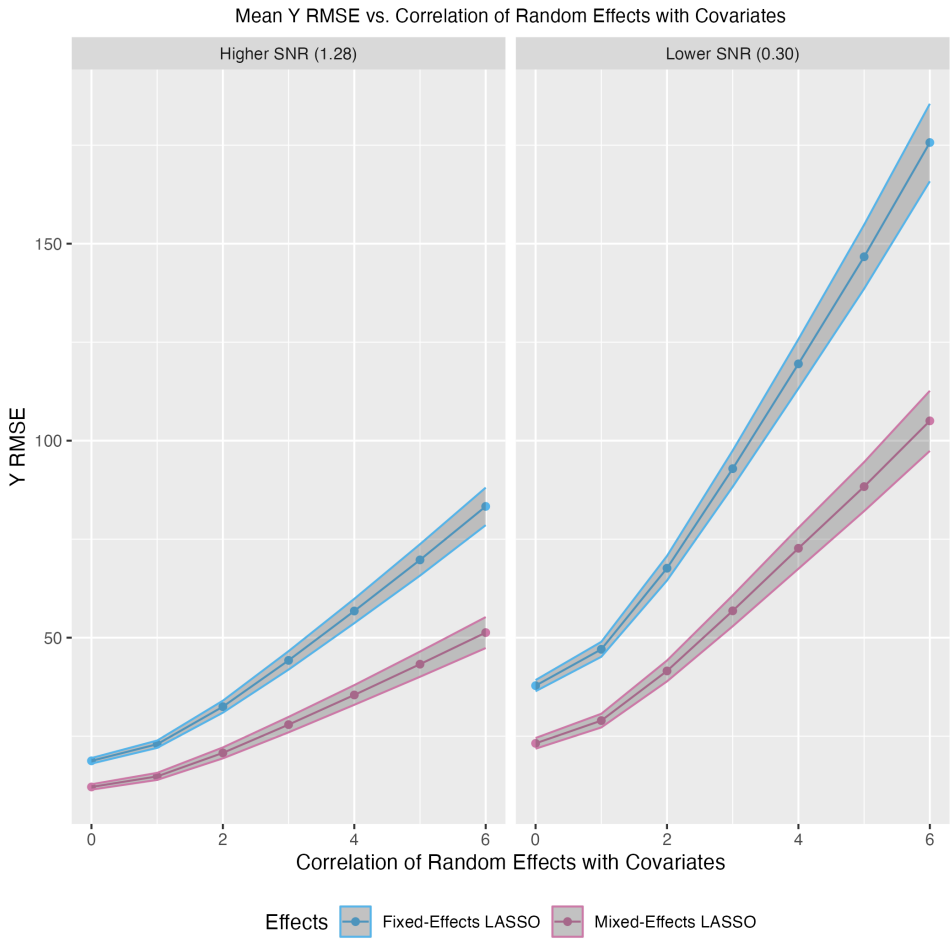

Figure A.1.15: Mean prediction RMSE of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation.

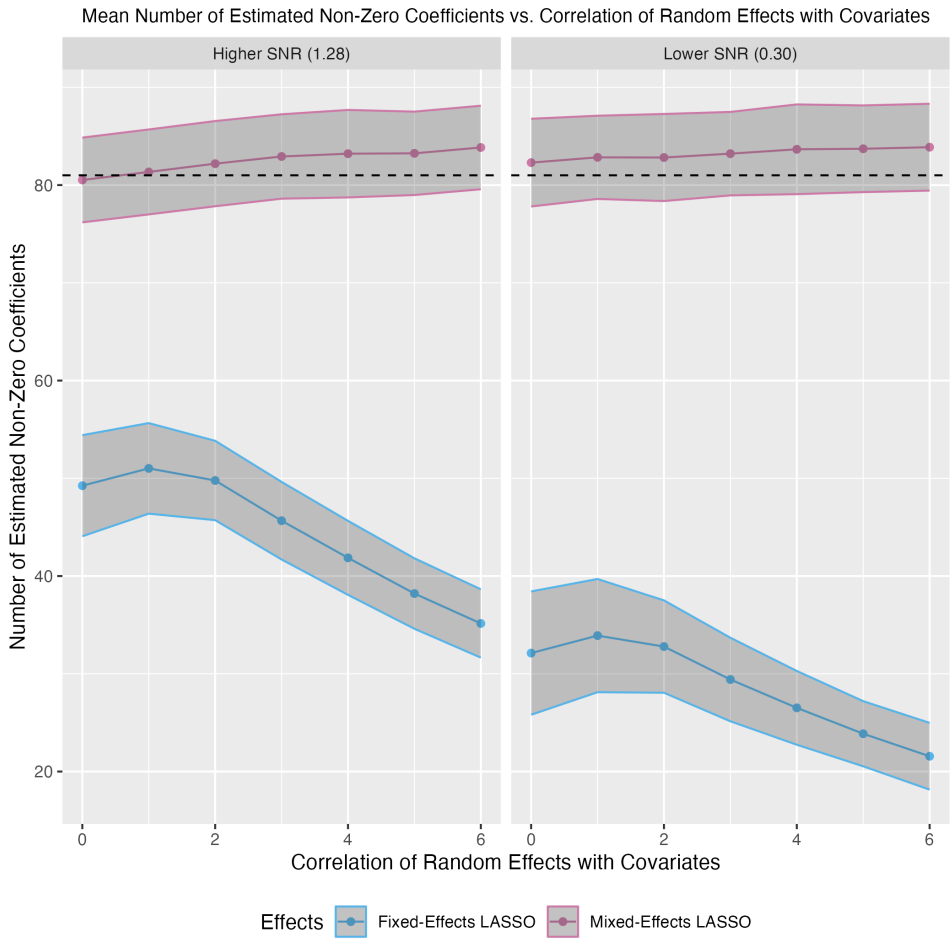

Figure A.1.16: Mean number of estimated non-zero coefficients of the fixed and mixed-effects LASSO, in relation to the correlation between the random effects and the covariance generation. The true value of non-zero coefficients is $w = 81$, denoted by the dotted line.

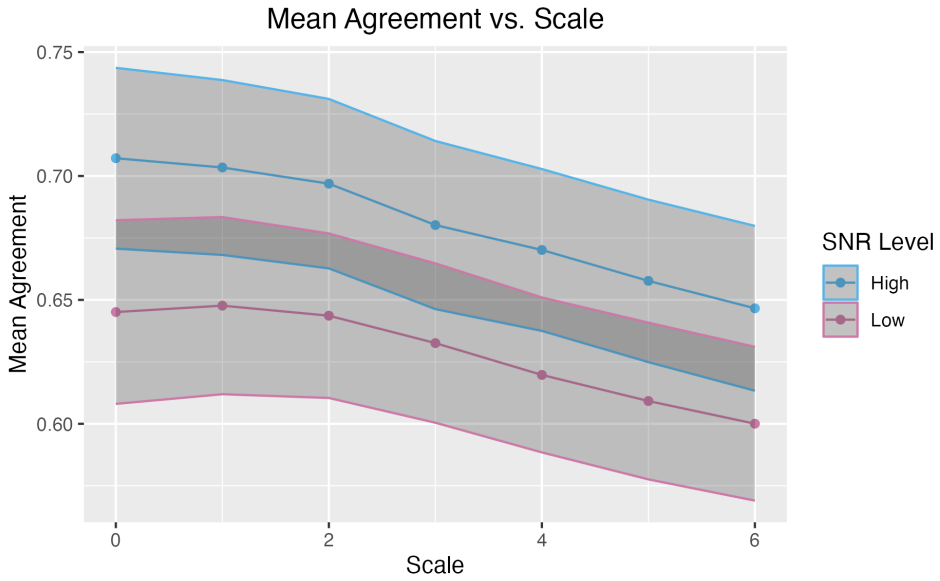

Figure A.1.17: Mean agreement between fixed and mixed-effects LASSO, in relation to the "scale" of the correlation between the random intercept and the covariate means.

A.1.4 Data Results

Below are the variables that were selected by both the fixed and mixed-effects LASSO, for all three outcomes.[1] As a reminder, these predictors all relate to the proportion of the time that a child or teacher engages in a specific action, since we averaged dummy variables across sweeps. For example, `o_c_towhom_Child_C.Oneotherchild` is the **proportion** of the time the child was observed to be talking to one other child. (We do not write "proportion of time that child was observed ..." in each bullet point for brevity.) All descriptions are taken from the COP and TOP manuals [35, 36].

For the MEFS Z-Score:

[1] We only include the intersection of the fixed and mixed-effects LASSO, since the mixed-effects LASSO selected well over 100 variables for each outcome.

- `o_c_towhom_Child_C.Oneotherchild`: Child was talking or listening to one other child.
- `o_c_interaction_Alone_AL`: Child was working alone in an activity unique from the activities of all others in the classroom (e.g., a child was using writing tools while other children were working on puzzles).
- `o_c_interaction_Unoccupied_UN`: Child was not attending to any particular learning-related activity.
- `o_c_typetask_None_N`: Child was not directly engaged in an activity or with materials, and is also not engaged in social talk.
- `o_c_focus_Literacy.Writing_LW`: Child was engaged in an activity that focused on writing and literacy.
- `o_c_involvement_MediumLow_ML_classmean`: The mean time the child's classmates were engaging and focusing in the learning activity at a low to medium level.
- `o_c_schedule_Playground_P_classsd`: The standard deviation of the proportion of sweeps that the classmates were on the playground.
- `o_c_schedule_WholeGroup_WG_classsd`: The standard deviation of the time that the child's classmates were in a setting with the whole group.
- `o_t_whom_o_SGT`: Teacher was talking or listening to a group of children (more than one child but less than 75% of the class) and one other teacher or assistant.
- `o_t_tone_o_N`: Teacher was looking displeased and exhibited annoyance or disappointment.

For the Leiter Test Score:

- `c_m8_goal_5_1`: One of the "magic 8" goals that are regarded to be consistently associated with residualized gain scores: this is the 5th goal, which is the proportion of time spent providing more sequential activities [44].
- `o_c_verbal_Fuss.Cry_FC`: Child was fussing, whining, yelling, crying, or arguing.
- `o_c_verbal_TalkSounds_TS`: Child was making noises, such as animal sounds.
- `o_c_towhom_WholeGroupNoTeacher_WG`: The child was talking or listening to 75% or more of the class, without the teacher present.
- `o_c_schedule_MealTime_MT`: Child was eating breakfast, lunch, or snacks.
- `o_c_focus_Math_M`: Child was engaged in an activity that focused on math.
- `o_c_towhom_WholeGroupNoTeacher_WG_classmean`: The average time the child's classmates spent talking or listening to 75% or more of the class, without the teacher present.
- `o_c_involvement_Medium_M_classmean`: The standard deviation of the time that the child's classmates spent engaging and focusing in the learning activity at a medium level.
- `o_c_verbal_Talk_T_classsd`: The standard deviation of the time that the child's classmates were talking in English.

- `o_c_schedule_Transition_T_classsd`: The standard deviation of the proportion of sweeps that the classmates were in a transition period.
- `o_c_typetask_Sequential_SQ_classsd`: The standard deviation of the time that the child's classmates were involved with activities or materials that involve a sequence of steps.
- `o_c_involvement_MediumHigh_MH_classsd`: The standard deviation of the time that the child's classmates spent engaging and focusing in the learning activity at a medium to high level.
- `o_c_involvement_MediumLow_ML_classsd`: The standard deviation of the time that the child's classmates spent engaging and focusing in the learning activity at a low to medium level.
- `o_t_whom_o_WG`: Teacher was talking or listening to at least 75% of the class.
- `o_t_task_o_MA`: Teacher was actively engaged in an activity required to run a classroom.
- `o_t_instruct_3`: Teacher was interacting with children using some open-ended questions.
- `o_t_focus_o_A`: Teacher was engaging in a learning activity that focused on art.
- `o_t_focus_o_LA`: Teacher was engaging in a learning activity that focused on letter sounds, names, and/or spelling.
- `o_t_focus_o_LW`: Teacher was engaging in a learning activity that focused on literacy and writing.
- `o_t_tone_o_N`: Teacher was looking displeased and exhibited annoyance or disappointment.

- `o_t_es_o_N`: Teacher was not exhibiting any behavior related to the emotions of the children.

The fixed-effects LASSO for the PCA-derived outcome shrank all coefficients to zero, so there is no intersection of predictors to list.

References

[1] Bernadette Daelmans, Gary L Darmstadt, Joan Lombardi, Maureen M Black, Pia R Britto, Stephen Lye, Tarun Dua, Zulfiqar A Bhutta, and Linda M Richter. Early childhood development: the foundation of sustainable development. *The Lancet*, 389(10064):9–11, 2017.

[2] Ashley Brunsek, Michal Perlman, Olesya Falenchuk, Evelyn McMullen, Brooke Fletcher, and Prakesh S Shah. The relationship between the early childhood environment rating scale and its revised form and child outcomes: A systematic review and meta-analysis. *PloS one*, 12(6):e0178512, 2017.

[3] Emily C. Hanno and Kathryn E. Gonzalez. The effects of teacher professional development on children's attendance in preschool. *Journal of Research on Educational Effectiveness*, 13(1):3–28, 2020.

[4] Tran D Keys, George Farkas, Margaret R Burchinal, Greg J Duncan, Deborah L Vandell, Weilin Li, Erik A Ruzek, and Carollee Howes. Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child development*, 84(4):1171–1190, 2013.

[5] Abbie Raikes, Natalie Koziol, Magdalena Janus, Linda Platas,

Tara Weatherholt, Anna Smeby, and Rebecca Sayre. Examination of school readiness constructs in tanzania: Psychometric evaluation of the melqo scales. *Journal of Applied Developmental Psychology*, 62:122–134, 2019.

[6] Christina Weiland, Kchersti Ulvestad, Jason Sachs, and Hirokazu Yoshikawa. Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, 28(2):199–209, 2013.

[7] M Ramaswami and R Bhaskaran. A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*, 2009.

[8] W Holmes Finch, Jocelyn E Bolin, and Ken Kelley. *Multilevel modeling using R*. Crc Press, 2019.

[9] Robert F Dedrick, John M Ferron, Melinda R Hess, Kristine Y Hogarty, Jeffrey D Kromrey, Thomas R Lang, John D Niles, and Reginald S Lee. Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1):69–102, 2009.

[10] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[11] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111, 2010.

[12] Stephen W Raudenbush and Anthony S Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage, 2002.

[13] D Betsy McCoach and Jill L Adelson. Dealing with dependence (part i): Understanding the effects of clustered data. *Gifted Child Quarterly*, 54(2):152–155, 2010.

[14] Ann A O'Connell and D Betsy McCoach. *Multilevel modeling of educational data*. IAP, 2008.

[15] Howard D Bondell, Arun Krishna, and Sujit K Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

[16] Alastair H Leyland and Peter P Groenewegen. *Multilevel modelling for public health and health services research: health in context*. Springer Nature, 2020.

[17] Song S Qian, Thomas F Cuffney, Ibrahim Alameddine, Gerard McMahon, and Kenneth H Reckhow. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, 91(2):355–361, 2010.

[18] W Holmes Finch et al. Modeling high dimensional multilevel data using the lasso estimator: A simulation study. *Journal of Statistical and Econometric Methods*, 7(1):51–75, 2018.

[19] Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, 24(2):137–154, 2014.

[20] Joop Hox. *Multilevel analysis techniques and applications*. Crc Press, 2002.

[21] Jürg Schelldorfer, Peter Bühlmann, and SARA VAN DE GEER. Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

[22] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[23] Tibshirani R. Friedman J. H. Hastie, T. *The elements of statistical learning: data mining, inference, and prediction, Second Edition*. Springer, 2009.

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

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

[26] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[27] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[28] Jonas Ranstam and JA Cook. Lasso regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.

[29] Edward Cripps, Chris Carter, and Robert Kohn. Variable selection and covariance selection in multivariate regression models. In D.K. Dey and C.R. Rao, editors, *Bayesian Thinking*, volume 25 of *Handbook of Statistics*, pages 519–552. Elsevier, 2005.

[30] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

[31] Chao Bian, Yawen Zhou, and Chao Qian. Robust subset selection by greedy and evolutionary pareto optimization. *arXiv preprint arXiv:2205.01415*, 2022.

[32] Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

[33] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

[34] Early learning study at harvard. https://zaentz.gse.harvard.edu/what-we-do/early-learning-study-harvard/. Accessed: 2022-08-31.

[35] Dale Farran and Karen Anthony. *Children Observation in Preschools (COP) Manual*. Peabody Research Institute.

[36] Dale Farran and Karen Anthony. *Teacher Observation in Preschools (TOP) Manual*. Peabody Research Institute.

[37] Yanming Li, Bin Nan, and Ji Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363, 2015.

[38] Nina Dörnemann and Holger Dette. Fluctuations of the diagonal entries of a large sample precision matrix. *arXiv preprint arXiv:2211.00474*, 2022.

[39] Cristan Farmer. *Leiter International Performance Scale-Revised (Leiter-R)*, pages 1732–1735. Springer New York, New York, NY, 2013.

[40] Peter Baggetta and Patricia A Alexander. Conceptualization and operationalization of executive function. *Mind, Brain, and Education*, 10(1):10–33, 2016.

[41] Designing monte carlo simulations in r. https://jepusto.github.io/Designing-Simulations-in-R/. Accessed: 2022-10-31.

[42] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. 2016.

[43] Artuur M Leeuwenberg, Maarten van Smeden, Johannes A Langendijk, Arjen van der Schaaf, Murielle E Mauer, Karel GM Moons, Johannes B Reitsma, and Ewoud Schuit. Comparing methods addressing multi-collinearity when developing prediction models. *arXiv preprint arXiv:2101.01603*, 2021.

[44] Dale C Farran, Deanna Meador, Caroline Christopher, Kimberly T Nesbitt, and Laura E Bilbrey. Data-driven improvement in prekindergarten classrooms: Report from a partnership in an urban district. *Child Development*, 88(5):1466–1479, 2017.