University of Colorado Anschutz Medical Campus

Project 4, Report

Joseph Froelicher

December 17, 2021

Contents

Introduction

The purpose of this analysis is to investigate several common variable selection techniques used primarily in linear regression. We are most interested in comparing the reliability of these variable selection methods when it comes to several measures of precision and accuracy, which will be detailed below. Most importantly, we want to know if variable selection techniques can pick out variables that are already known to be significant. And how that relates to the effect size of the variable coefficients. Additionally, we are also interested in how well variable selection techniques ignore variables that are not statistically significant.

To investigate variable selection techniques, the analysis team will perform a simulation of data. There are two primary simulations for this analysis. The first is to simulate data using a sample size of 250 and the second is to simulated data using a sample size of 500. For each sample size, we will simulate both correlated and uncorrelated data. The details of those simulations are outlined below.

It is hypothesized that all variable selection methods will select variables with larger effect sizes more accurately than those with small effect sizes. It is also apparent that more conservative selection methods will wrongly select insignificant variables less frequently that the more anti-conservative selection techniques. Lastly, it is hypothesized that methods that use penalized regression techniques will select insignificant variables at a much higher rate than other selection methods.

Materials and Methods

Variable Selection

This analysis will explore seven variable selection techniques, from two general selection fields, backwards selection, and penalized regression.

Backwards Selection

Backwards selection is the older more traditional variable selection method. It is done by fitting a model with all available variables, and iteratively removing insignificant predictors one variable at a time based on certain criteria. The three criteria we are using for this analysis, are Likelihood Ratio Test (LRT) p-values, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The formulas for each are provided below.

$$\lambda_{LR} = -2\big[l(\theta_0) - l(\hat{\theta})\big] \tag{1}$$

$$AIC = -2k - 2l(\hat{\theta}) \tag{2}$$

$$BIC = -kln(n) - 2l(\hat{\theta}) \tag{3}$$

$l(\theta)$: log-likelihood

n : sample size

k : number of parameters

Penalized Regression

We are also interested in two variations of two penalized regression model selection techniques. The two methods are LASSO, and Elastic Net. We will use both a fixed value LASSO and Elastic Net, and 10-fold cross-validated LASSO and Elastic Net. This means that tuning parameters are either prespecified (fixed), or cross-validated (CV). These methods find the best fit model by optimized the log-likelihood plus a penalty term, which is some non-negative function of the all parameters in the model. The penalty terms are as follows.

LASSO:

$$\lambda||\theta||_1 \tag{4}$$

Elastic Net:

$$\lambda(\alpha||\theta|_1 + (1-\alpha)||\theta||_2) \tag{5}$$

λ, α : tuning parameters

It should be noted that to fit final models after using penalized regression for variable selection, standard linear regression modeling was used. This was only to obtain estimates for model comparison (see limitations).

Simulation

Model

The standard linear regression model is as follows:

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ji} + \epsilon_i \tag{6}$$

$$X_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$i = 1, ..., n$$

$$j = 1, ..., p$$

The coefficients from which the data were simulated for each scenario are as follows:

$$\boldsymbol{\beta} = \left[\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right]$$

The each of the four scenarios mentioned previously use the information provided above. Additionally, scenarios *1B* and *2B* induce a correlation of $\rho = 0.4$ for the predictor variables.

Evaluation

For each aim, 1000 data sets were simulated and then analyzed for eight evaluation parameters of interest. The evaluation parameters are provided below.

$$\text{True Positive Rate } (TPR) = \frac{\text{number of times } x_i \text{ is selected}}{1000}, \text{for each } \beta_i \neq 0 \tag{7}$$

$$\text{False Positive Rate } (FPR) = \frac{\text{number of times } x_i \text{ is selected}}{15 * 1000}, \ \forall \beta_i = 0 \tag{8}$$

$$\text{False Discovery Rate } (FDR) = \frac{\text{number of } x_i \text{ w}/x_i = 0 \text{ that are selected}}{\text{number of } x_i \text{that are selected}}, \text{across all 1000 simulations} \tag{9}$$

$$\text{Average Bias } (\text{bias}_0) = \frac{1}{n_\beta} \sum_{\forall \beta} \beta_{observed} - \beta_{expected}, \ \forall \beta_i = 0 \tag{10}$$

$$\text{Bias } = \beta_{observed} - \beta_{expected}, \text{for each} \beta_i \neq 0 \tag{11}$$

$$\text{Coverage } = \frac{\text{number of times the 95\% CI for } \hat{\beta}_i \text{ contains } x_i}{\text{number of times } x_i \text{ is selected}}, \text{for each } \beta_i \neq 0 \tag{12}$$

$$\text{Type I error rate } = \frac{\text{number of times } x_i \text{ selected and p-value } < 0.05}{\text{number of times } x_i \text{ is selected}}, \forall \beta_i = 0 \tag{13}$$

$$\text{Type II error rate } = \frac{\text{number of times } x_i \text{ is selected and p } \geq 0.05}{\text{number of times } x_i \text{ is selected}}, \text{ for each} \beta_i \neq 0 \tag{14}$$

We expect type I error rates and type II error rates to be inflated and dflated respectively, due to them being conditioned on the number of times a variable is selected. All statistics presented above are contained between 0 and 1. A favorable value for TPR and coverage rate is 1, and a favorable value for FPR, FDR, bias, type I error, average bias, and type II error is 0.

Results

Bias

Figure 1 shows the bias for each method, with each of variables 1 through 5 as different colors, stratified by aim. Bias was fairly small, and distributed randomly for scenario 1A. Bias was consistently higher for all methods in aim's 1A and 2a than it was in aim's 1B and 2B. Bias was reliably high for all aims after aim 1A using fixed LASSO penalized regression. Bias was reliably low for all aims after aim 1A using the LRT p-value variable selection method.

\subsection{95% Confidence Interval Coverage}

The coverage of the 95% confidence intervals is shown in figure 2. As can be seen visually, there was fairly consistent coverage across all aims and all selection techniques. All model selection techniques had more trouble on average capturing variable $X1$. For the independent data simulations (1A & 2A), fixed LASSO penealized regression was particularly bad at capturing variable $X1$.

True Positive Rate

Presented in figure 3 is the true positive rate for each variable and each method, stratified by aim. Figure 3 shows very clearly that that variable $X1$ was selected correctly at a much lower rate than variables $X2 - X5$. And the true positive rate was particularly low for fixed LASSO penalized regression and small sample sizes (aims 1A & 2A).

Average False Discovery Rate

Presented in figure 4 is the average FDR across variables $X6 - X20$. There are two things of note here, the first is that BIC backward selection had the lowest FDR consistently across all aims. The second is that the cross-validated penalized regression methods had particularly high FDR across all aims.

False Positive Rate

Similar to the FDR, the average FPR across variables $X6 - X20$, shown in figure 5 was consistently lowest across all aims using the BIC backward selection method. And highest across all aims using the cross-validate penalized regression methods.

Type I error Rate

Shown in figure 6 is the average type I error across variables $X6 - X20$. The type I error is consistently high aross all aims using the BIC backward selection method. The type I error rate was also notably high in the fixed penalized regression selection methods for independent data (aims 1A and 2A).

Conclusion

Discussion

This simulation demonstrated some interesting findings. First and foremost, in response to the hypotheses, variable selection on data simulated for larger data sets did seem to be slightly better (aims 2A & 2B) than those of smaller sample size (aims 1A & 1B). It was also confirmed that of the non-zero coefficients ($\beta_1 - \beta_5$) the smallest effects size (β_1) was selected less frequently by all selection methods than the other four variables. Additionally, the BIC is considered to be conservative (has a significant penalty, particularly for large sample sizes), and this selection did in fact wrongly select insignificant variables at a much lower rate than any other method. Lastly, the cross-validated penalized regression models did seem to be selecting insignificant variables at a much higher rate than any other model on average across all aims. This is likely due to the design of these methods, which are designed to optimized prediction, so they favor leaving in variables that explain small amounts of variability in the outcome.

Limitations

One significant limitation for this analysis was the way that regression models were built after selection using all of the penalized regression methods. After all four of these models selected significant variables, and then linear regression models were reran to evaluate coefficient values, 95% confidence intervals, and p-values. Additionally, it would be of interest to evaluate these methods on much smaller (< 50) and much larger (> 1000) data in addition to the sample sized used in this analysis, this may teach us more about varaible selection in extreme circumstances.

Reproducibility

Each of the analyses for all aims were performed in `R 4.1.2`. The simulated data was generated using `hdrm::genData()`.

Code for this analysis is available from `https://github.com/BIOS6624-UCD/bios6624-JoeFroelicher/project4`,

under the branch `project4`. Special note `@ehccooper` for assistance on both report writing and mathematical formulation.

Figures

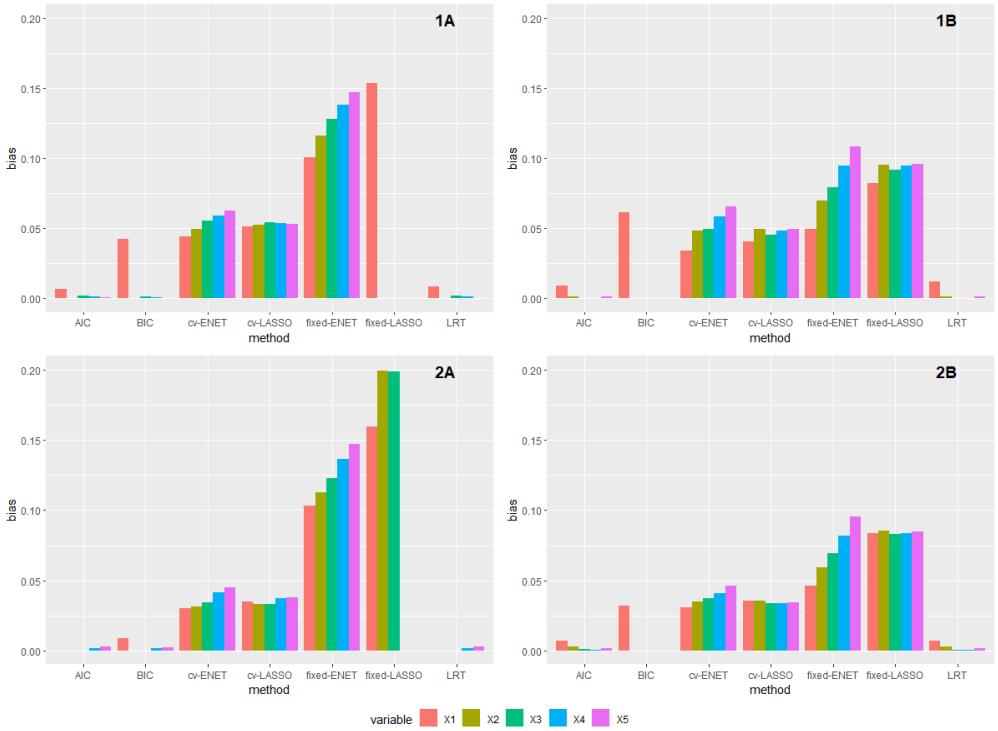

Figure 1: Bias for each of variables X1-X5, stratified by aim.

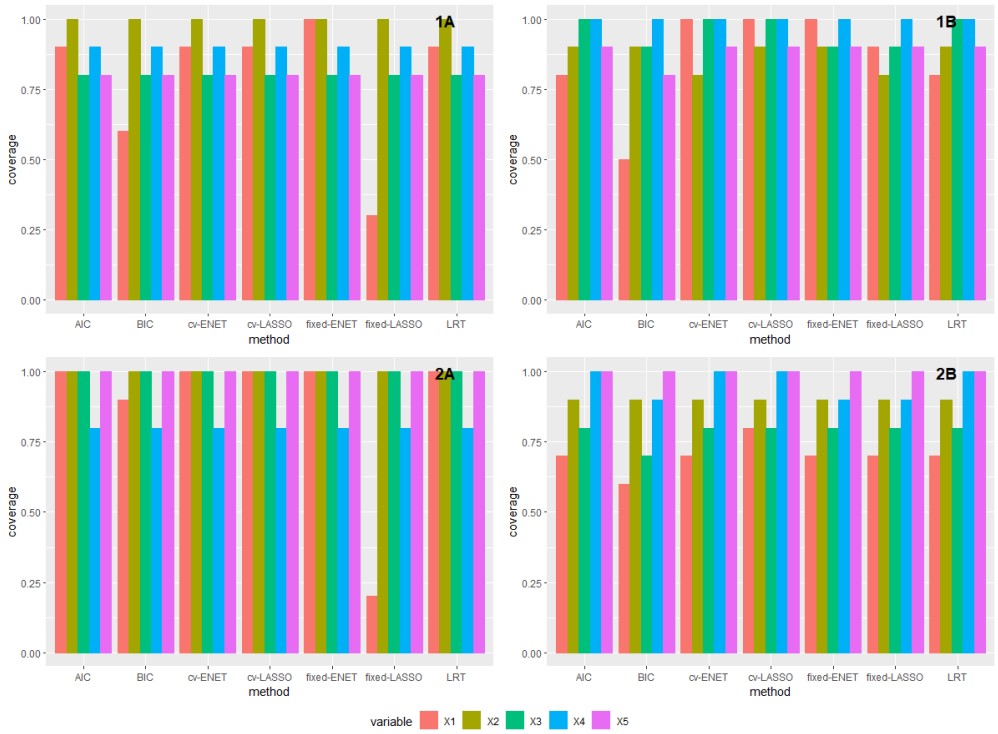

Figure 2: 95% Confidence interval coverage rate, for each of variables X1-X5, stratified by aim.

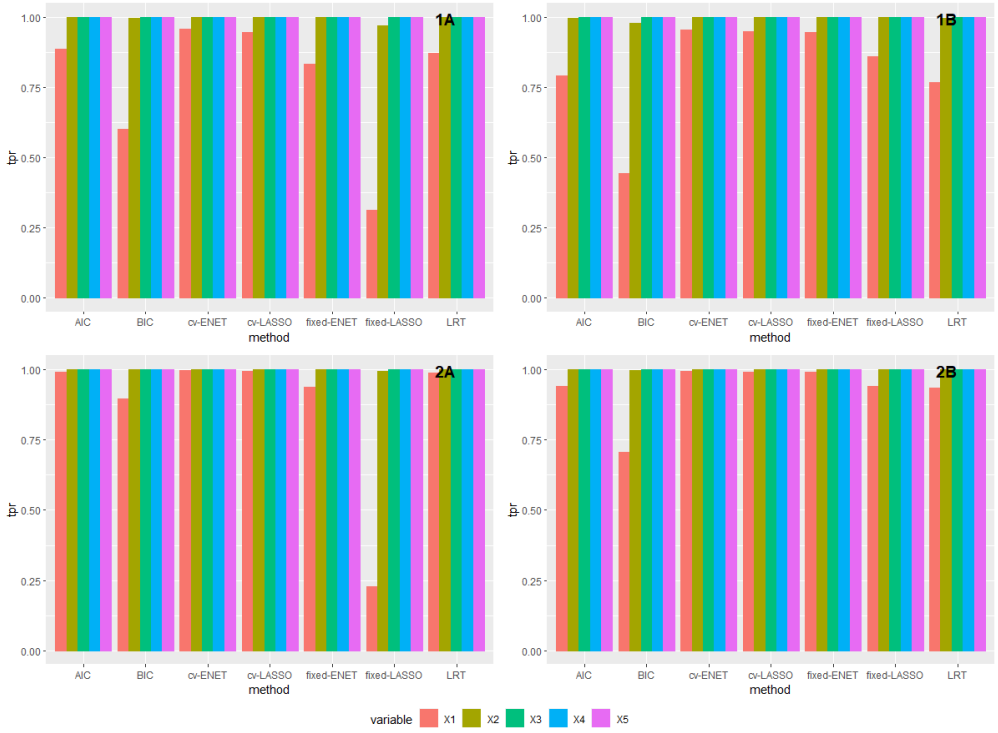

Figure 3: True positive rate for each of variables X1-X5, stratified by aim.

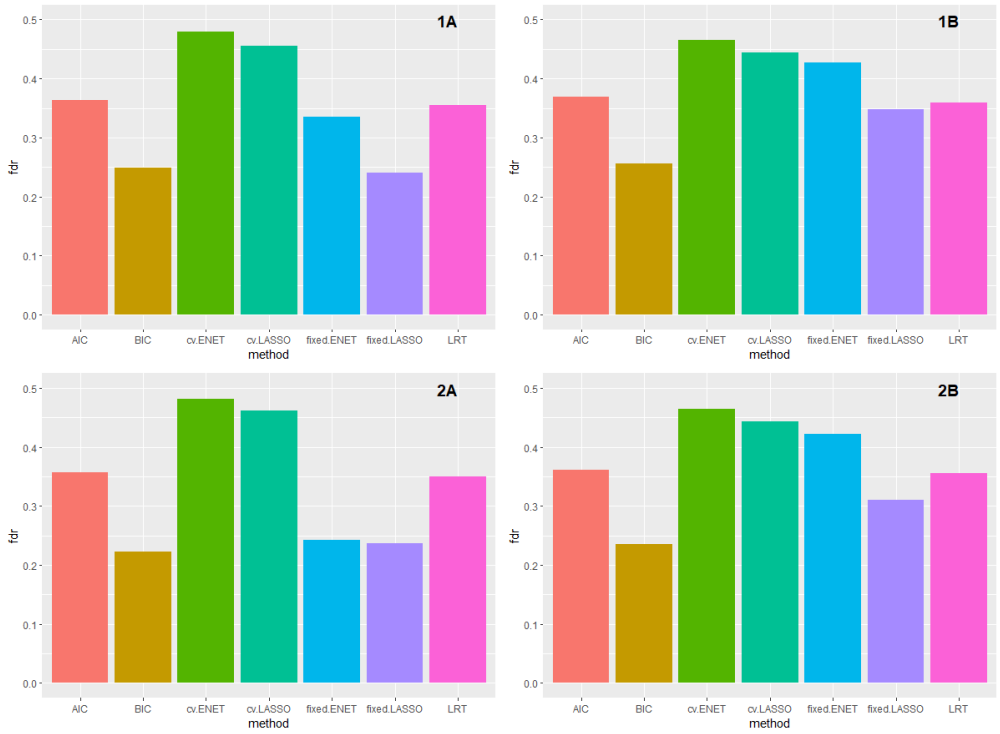

Figure 4: Average false discovery rate of variables X6-X20, stratified by aim.

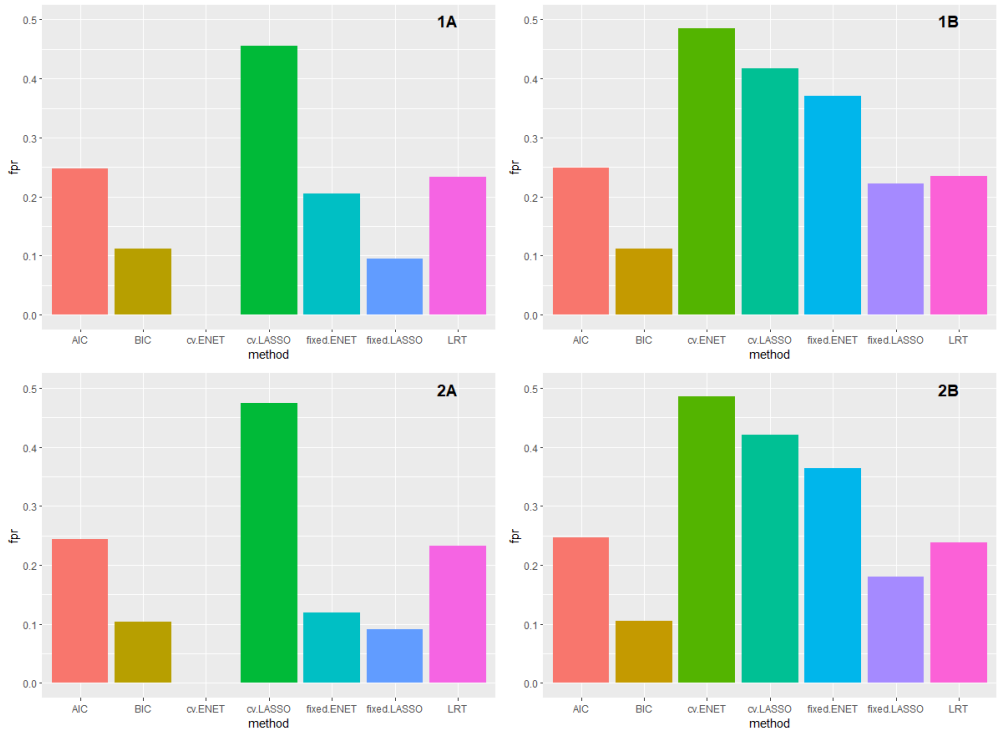

Figure 5: Average false positive rate for variables X6-X20, stratified by aim.

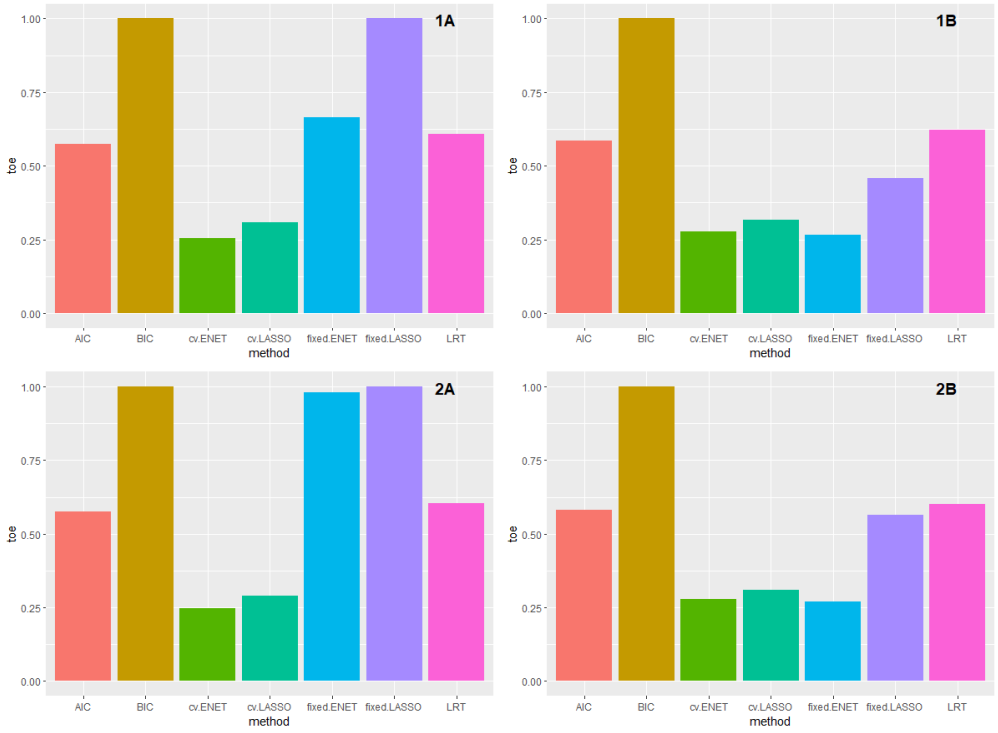

Figure 6: Average type I error rate for variables X6-X20, stratified by aim.

•

Supplementary Figures

	LRT	AIC	BIC	cv.LASSO	fixed.LASSO	cv.ENET	fixed.ENET
FDR	0.3548119469	0.3634666667	2.485303e-01	0.45485102	0.23968656	0.47891615	0.33586197
FPR	0.2332727273	0.2478181818	1.114545e-01	0.45518182	0.09454545	0.54309091	0.20527273
TPR: X1	0.8700000000	0.8870000000	6.000000e-01	0.94500000	0.31300000	0.95700000	0.83200000
TPR: X2	1.0000000000	1.0000000000	9.960000e-01	1.00000000	0.96800000	1.00000000	1.00000000
TPR: X3	1.0000000000	1.0000000000	1.000000e+00	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X4	1.0000000000	1.0000000000	1.000000e+00	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X5	1.0000000000	1.0000000000	1.000000e+00	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X1	0.9000000000	0.9000000000	6.000000e-01	0.90000000	0.30000000	0.90000000	1.00000000
CI Coverage: X2	1.0000000000	1.0000000000	1.000000e+00	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X3	0.8000000000	0.8000000000	8.000000e-01	0.80000000	0.80000000	0.80000000	0.80000000
CI Coverage: X4	0.9000000000	0.9000000000	9.000000e-01	0.90000000	0.90000000	0.90000000	0.90000000
CI Coverage: X5	0.8000000000	0.8000000000	8.000000e-01	0.80000000	0.80000000	0.80000000	0.80000000
Type II Error: X1	0.1505747126	0.1691093574	0.000000e+00	0.23809524	0.00000000	0.25287356	0.12740385
Type II Error: X2	0.0010000000	0.0010000000	0.000000e+00	0.00100000	0.00000000	0.00100000	0.00100000
Type II Error: X3	0.0000000000	0.0000000000	0.000000e+00	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X4	0.0000000000	0.0000000000	0.000000e+00	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X5	0.0000000000	0.0000000000	0.000000e+00	0.00000000	0.00000000	0.00000000	0.00000000
Type I Error: X6-X20	0.6091192518	0.6091192518	6.091193e-01	0.60911925	0.60911925	0.60911925	0.60911925
Bias: X1	0.0081810527	0.0066528079	4.238843e-02	0.05084773	0.15359675	0.04412633	0.10054577
Bias: X2	-0.0002979288	-0.0003007770	9.311912e-05	0.05245954	0.20201193	0.04931785	0.11611579
Bias: X3	0.0015248882	0.0015505329	1.108491e-03	0.05401982	0.20508868	0.05505351	0.12803322
Bias: X4	0.0014408893	0.0014232443	6.936390e-04	0.05372633	0.20456557	0.05908678	0.13773248
Bias: X5	0.0001894812	0.0003705743	-1.814918e-04	0.05287503	0.20374602	0.06260326	0.14702667
Bias: X6-X20	0.0758397593	0.0759465940	7.600946e-02	0.07082382	0.05724439	0.06992601	0.06252535

Figure 7: Table of results for aim 1a.

	LRT	AIC	BIC	cv.LASSO	fixed.LASSO	cv.ENET	fixed.ENET
FDR	0.3595364423	0.3684635032	0.256388773	0.44422928	0.34768354	0.46502309	0.42746331
FPR	0.2340909091	0.2491818182	0.111272727	0.41709091	0.22172727	0.48527273	0.37072727
TPR: X1	0.7670000000	0.7910000000	0.444000000	0.94700000	0.86000000	0.95500000	0.94600000
TPR: X2	0.9950000000	0.9950000000	0.977000000	1.00000000	0.99800000	1.00000000	1.00000000
TPR: X3	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X4	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X5	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X1	0.8000000000	0.8000000000	0.500000000	1.00000000	0.90000000	1.00000000	1.00000000
CI Coverage: X2	0.9000000000	0.9000000000	0.900000000	0.90000000	0.80000000	0.80000000	0.90000000
CI Coverage: X3	1.0000000000	1.0000000000	0.900000000	1.00000000	0.90000000	1.00000000	0.90000000
CI Coverage: X4	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X5	0.9000000000	0.9000000000	0.800000000	0.90000000	0.90000000	0.90000000	0.90000000
Type II Error: X1	0.2307692308	0.2503160556	0.000000000	0.52375924	0.49069767	0.52460733	0.59725159
Type II Error: X2	0.0050251256	0.0060301508	0.000000000	0.03300000	0.03507014	0.03500000	0.04600000
Type II Error: X3	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X4	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X5	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type I Error: X6-X20	0.6217475728	0.6217475728	0.621747573	0.62174757	0.62174757	0.62174757	0.62174757
Bias: X1	0.0115956964	0.0086924618	0.061422954	0.04048961	0.08199685	0.03404516	0.04918397
Bias: X2	0.0010868476	0.0013119370	-0.002195771	0.04949601	0.09536246	0.04805946	0.06980785
Bias: X3	-0.0038997810	-0.0038385671	-0.010941092	0.04508399	0.09160583	0.04932311	0.07881023
Bias: X4	-0.0007218429	-0.0007097277	-0.008048071	0.04810259	0.09473230	0.05842638	0.09443131
Bias: X5	0.0010148084	0.0009097838	-0.005671422	0.04957292	0.09574457	0.06559100	0.10816670
Bias: X6-X20	0.0763366583	0.0762346917	0.077677935	0.08147893	0.07248254	0.08168769	0.07884986

Figure 8: Table of results for aim 1b.

	LRT	AIC	BIC	cv.LASSO	fixed.LASSO	cv.ENET	fixed.ENET
FDR	0.3595364423	0.3684635032	0.256388773	0.44422928	0.34768354	0.46502309	0.42746331
FPR	0.2340909091	0.2491818182	0.111272727	0.41709091	0.22172727	0.48527273	0.37072727
TPR: X1	0.7670000000	0.7910000000	0.444000000	0.94700000	0.86000000	0.95500000	0.94600000
TPR: X2	0.9950000000	0.9950000000	0.977000000	1.00000000	0.99800000	1.00000000	1.00000000
TPR: X3	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X4	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X5	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X1	0.8000000000	0.8000000000	0.500000000	1.00000000	0.90000000	1.00000000	1.00000000
CI Coverage: X2	0.9000000000	0.9000000000	0.900000000	0.90000000	0.80000000	0.80000000	0.90000000
CI Coverage: X3	1.0000000000	1.0000000000	0.900000000	1.00000000	0.90000000	1.00000000	0.90000000
CI Coverage: X4	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X5	0.9000000000	0.9000000000	0.800000000	0.90000000	0.90000000	0.90000000	0.90000000
Type II Error: X1	0.2307692308	0.2503160556	0.000000000	0.52375924	0.49069767	0.52460733	0.59725159
Type II Error: X2	0.0050251256	0.0060301508	0.000000000	0.03300000	0.03507014	0.03500000	0.04600000
Type II Error: X3	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X4	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X5	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type I Error: X6-X20	0.6217475728	0.6217475728	0.621747573	0.62174757	0.62174757	0.62174757	0.62174757
Bias: X1	0.0115956964	0.0086924618	0.061422954	0.04048961	0.08199685	0.03404516	0.04918397
Bias: X2	0.0010868476	0.0013119370	-0.002195771	0.04949601	0.09536246	0.04805946	0.06980785
Bias: X3	-0.0038997810	-0.0038385671	-0.010941092	0.04508399	0.09160583	0.04932311	0.07881023
Bias: X4	-0.0007218429	-0.0007097277	-0.008048071	0.04810259	0.09473230	0.05842638	0.09443131
Bias: X5	0.0010148084	0.0009097838	-0.005671422	0.04957292	0.09574457	0.06559100	0.10816670
Bias: X6-X20	0.0763366583	0.0762346917	0.077677935	0.08147893	0.07248254	0.08168769	0.07884986

Figure 9: Table of results for aim 2a.

	LRT	AIC	BIC	cv.LASSO	fixed.LASSO	cv.ENET	fixed.ENET
FDR	0.3556581986	0.3608562691	0.235056543	0.44386973	0.31020472	0.46488178	0.42216359
FPR	0.2380000000	0.2467272727	0.105818182	0.42127273	0.18045455	0.48618182	0.36363636
TPR: X1	0.9350000000	0.9410000000	0.707000000	0.99300000	0.94100000	0.99600000	0.99100000
TPR: X2	1.0000000000	1.0000000000	0.998000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X3	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X4	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
TPR: X5	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
CI Coverage: X1	0.7000000000	0.7000000000	0.600000000	0.80000000	0.70000000	0.70000000	0.70000000
CI Coverage: X2	0.9000000000	0.9000000000	0.900000000	0.90000000	0.90000000	0.90000000	0.90000000
CI Coverage: X3	0.8000000000	0.8000000000	0.700000000	0.80000000	0.80000000	0.80000000	0.80000000
CI Coverage: X4	1.0000000000	1.0000000000	0.900000000	1.00000000	0.90000000	1.00000000	0.90000000
CI Coverage: X5	1.0000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000
Type II Error: X1	0.1090909091	0.1158342189	0.000000000	0.25881168	0.19659936	0.26706827	0.30070636
Type II Error: X2	0.0010000000	0.0010000000	0.000000000	0.00200000	0.00200000	0.00200000	0.00200000
Type II Error: X3	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X4	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type II Error: X5	0.0000000000	0.0000000000	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000
Type I Error: X6-X20	0.5996944232	0.5996944232	0.599694423	0.59969442	0.59969442	0.59969442	0.59969442
Bias: X1	0.0071000500	0.0066195778	0.031825801	0.03557894	0.08370560	0.03068916	0.04643635
Bias: X2	0.0025798425	0.0025750702	-0.001426311	0.03571541	0.08549081	0.03483673	0.05900540
Bias: X3	0.0006446406	0.0007469691	-0.004322095	0.03345947	0.08291176	0.03695965	0.06934660
Bias: X4	0.0006284800	0.0006277327	-0.003793786	0.03351838	0.08335204	0.04102698	0.08197346
Bias: X5	0.0013885291	0.0015163981	-0.003468373	0.03437387	0.08457906	0.04600056	0.09525418
Bias: X6-X20	0.0761340599	0.0760813313	0.076465614	0.07992304	0.07030291	0.08014254	0.07603708

Figure 10: Table of results for aim 2b.

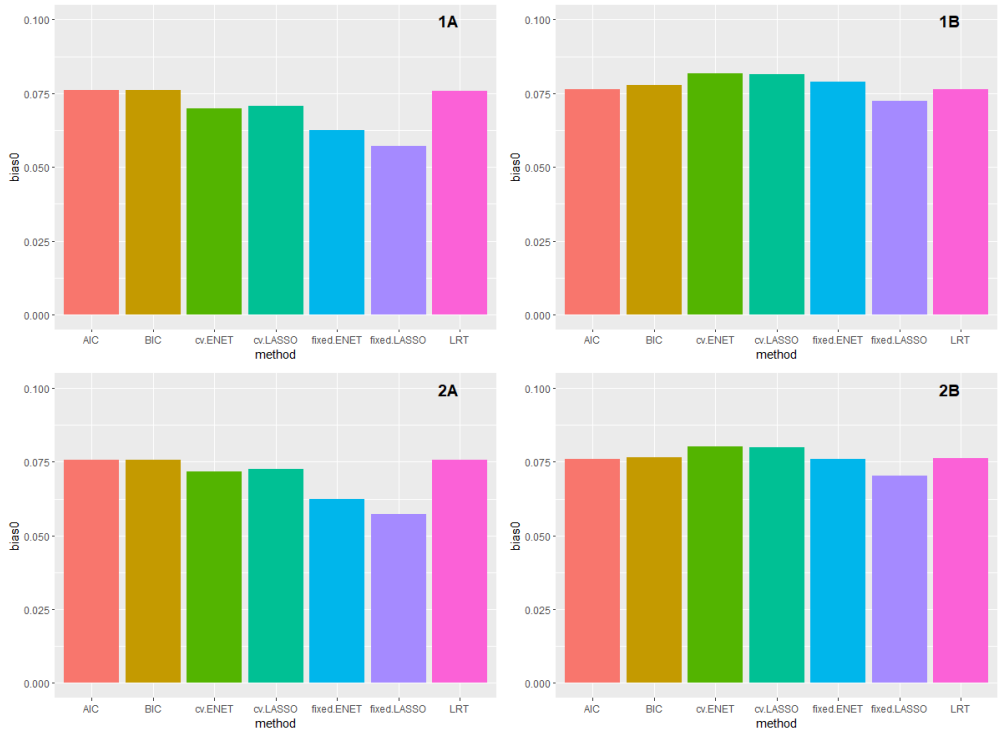

Figure 11: Average bias for variables X6-X20.

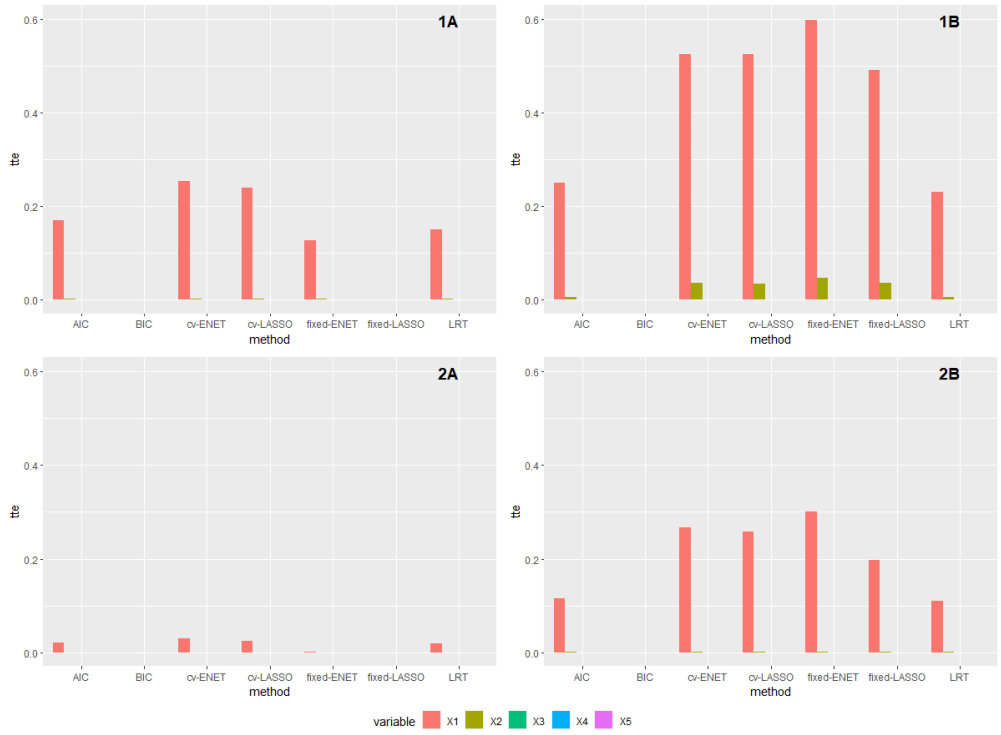

Figure 12: Type II error for each of variables X1-X5, stratified by aim.