Interpretation of Benchmark Results

# Benchmark 1: Least Absolute Shrinkage and Selection Operator

## True Positive Rates for BM1

The True Positive Rate (aka sensitivity) is the proportion of all structural variables (true factors) that the variable selection algorithm at hand selects for inclusion. Its range is . The TPR is the complement of the False Negative Rate, i.e., , thus, I won’t be including a section here interpreting LASSO’s False Negative Rate.

Remarkably, LASSO scores a perfect 100% on its TPR for all different levels of Collinearity, different number of candidate regressors which are structural, and different levels of error variance! Furthermore, even when breaking the results down further to the number of False Positives and Negatives for each of the 260,000 datasets, none of them have even a single False Negative. This is truly stellar performance on this metric, and much better than I expected to perform on this metric.

## True Negative Rates (aka Specificity)

The True Negative Rate is the proportion of all irrelevant predictor candidates, i.e., nonstructural variables which are correctly identified (so they are called True Negatives) as such by a given variable selection algorithm (e.g. LASSO). A high TNR indicates that the model is good at correctly identifying negatives, i.e., features that should not be selected in this case; in other words, it gives a sense of how good an algorithm does at avoiding Type-II errors. Its range is . The TNR is the complement of the False Positive Rate, i.e., , consequently, one only need analyze one of them. The overall mean TNR for all 260k datasets was 0.867, with 677,013 total False Positives being selected, which is another indication that this LASSO ran using the cv.glmnet() function from the glmnet package in R performed well.

I ran a multiple linear regression with mean TNR as the dependant variable , which is able to capture possible relationships with the degree of ‘Collinearity’ within a dataset (where 'Collinearity’ is the slope coefficient of a regressor as a function of X1 which is randomized as positive or negative), the number of structural variables in the population model, error variance, and for any possible interaction effects as well. Put formally, I ran:

Estimated equation:

So, there are statistically significant negative relationships for the TNR with Collinearity, the number of True Factors, and the interaction between those two, but not with the error variance.

Residuals:

Residual standard error: 0.01685 on 513 degrees of freedom

Multiple R-squared: 0.9578, Adjusted R-squared: 0.9573

F-statistic: 1939 on 6 and 513 DF, p-value: < 2.2e-16

A graph with a number of blue squares

Description automatically generated with medium confidence

#### **Summary for the LASSO’s True Negative Rate**

1. The True Negative Rates were extremely high across the board, this is a positive indicator of its performance.
2. Collinearity alone has a statistically significant impact on the True Negative Rate, and the sign on that impact is negative.
3. The number of true factors alone also has a statistically significant impact on the True Negative Rate, and once again, it is a negative impact.
4. There is a measurable (non-random) interaction effect on the TNR of the level of collinearity and the number of true factors separate. over and above their individual combined effects.
5. The level of error variance alone does not have any statistically measurable (non-random) impact on the True Negative Rate.

## Overspecified Models Selected by the 1st Benchmark

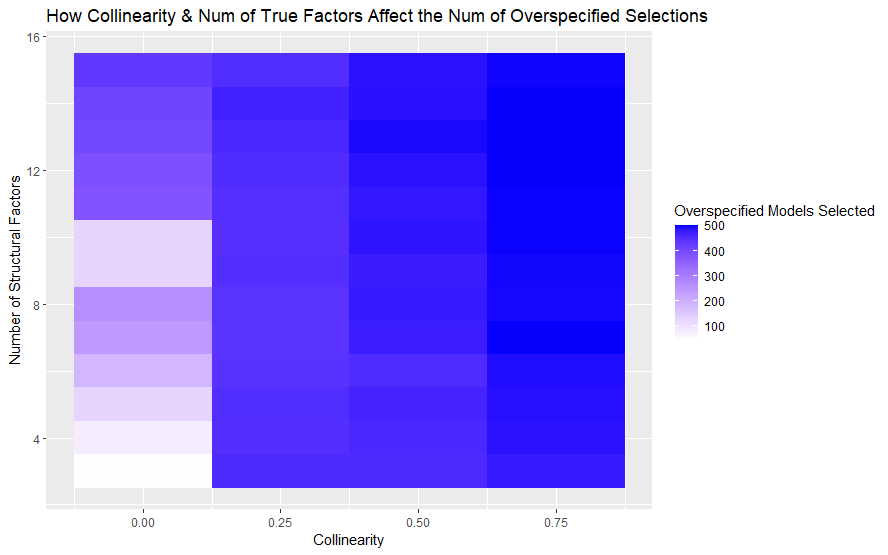
An overspecified regression (also known as an extraneous variable model) here is defined as one which includes all true factors, and at least one extraneous, i.e., extra factor (aka a False Positive) as well. Having a large number of overspecified models selected by a given variable selection algorithm informs you that the algorithm is not performing ideally because it is including lot of irrelevant predictors. In this way, it is similar to the False Positive Rate, and it’s complement, the True Negative Rate, but they are not the same thing. The total number of overspecified models selected by LASSO was 221,269.

Taking the same approach, I estimated the following equation:

Residual standard error: 65.6 on 513 degrees of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.69

F-statistic: 192.4 on 6 and 513 DF, p-value: < 2.2e-16



### The Conundrum of Very High TNRs (Good), Perfect TPRs (Great), and Also Very High Numbers of Overspecified Models (Bad)

How can there be a low False Positive Rate of 0.133 when the proportion of all selected models which are overspecified , should not those two at the very least be on the same side of the middle of their ranges, i.e., both be above or below 0.5?

The discrepancy is probably due to the fact that the True Positive Rate and the True Negative Rate are performance metrics that tell you how well selections were done at the level of each variable included in the selected model by the algorithm, whereas the number of overspecified models selected is a performance metric which is at the level of the overall regression model. So a 16 variable regression specification selected by LASSO on a dataset whose structural model is actually 15 variables will have a TNR of which is very high and quite good, but if evaluated by itself as a set of sample datasets size 1 rather than 260k, then it would have a proportion of all selected models which are overspecified of 100%.

Collinearity is also a major factor here, if there's collinearity among predictors, Lasso may include more than just the necessary predictors in the selected regression model. In such a scenario, it might still capture all the true predictors (hence TPR = 1), but also include extraneous variables that are correlated with the true predictors.

## Correctly Specified Models Selected

All of the ANOVA tests for the number of correctly specified models selected gave the same results as the ANOVA tests for the number of overspecified models selected. Because of this, I do not repeat the same tests for the number of correctly specified models for the 2nd and 3rd Benchmarks. The total number of correctly specified models selected by LASSO was 38,731.

Estimated equation:

Residual standard error: 65.6 on 513 degrees of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.69

F-statistic: 192.4 on 6 and 513 DF, p-value: < 2.2e-16

The number of correctly specified models selected has statistically significant negative relationships with Collinearity and the number of True Factors; but, plot twist, their interaction has a positive statistically significant effect on it.

A graph with purple squares

Description automatically generated with medium confidence

# Benchmark 2: Backward Elimination Stepwise Regression

## True Positive Rates for BM2

Like LASSO, there was a mean TPR of 1. I honestly don’t know what to make of this.

## True Negative Rates for BM2

The overall mean True Negative Rate for Backward Elimination was 0.799, and a total of 1,096,776 False Positives included.

Estimated equation:

Residual standard error: 0.006764 on 513 degrees of freedom

Multiple R-squared: 0.8887, Adjusted R-squared: 0.8874

F-statistic: 682.6 on 6 and 513 DF, p-value: < 2.2e-16

A graph of blue squares

Description automatically generated with medium confidence

## Overspecified Models Selected by the 2nd Benchmark

The total number of overspecified models selected by Backward Elimination was 256,090, and it selected 0 underspecified models, leaving just 3,910 correctly specified models selected.

Estimated equation:

Residual standard error: 4.195 on 513 degrees of freedom

Multiple R-squared: 0.6971, Adjusted R-squared: 0.6935

F-statistic: 196.7 on 6 and 513 DF, p-value: < 2.2e-16

A graph of blue squares

Description automatically generated with medium confidence

# Benchmark 3: Forward Selection Stepwise Regression

## True Positive Rates for BM3

Like LASSO and Backward Elimination, there was a mean TPR of 1.

## True Negative Rates for BM3

The overall mean True Negative Rate for Forward Selection was 0.823, and there are a total of 942,252 False Positives included.

Estimated equation:

Residual standard error: 0.006764 on 513 degrees of freedom

Multiple R-squared: 0.8887, Adjusted R-squared: 0.8874

F-statistic: 682.6 on 6 and 513 DF, p-value: < 2.2e-16

A graph of blue and purple squares

Description automatically generated

## Overspecified Models Selected by the 3rd Benchmark

The total number of overspecified models selected by Backward Elimination was 255,178, and it selected 0 underspecified models, leaving just 4,822 correctly specified models selected.

Estimated equation:

Residual standard error: 4.2 on 513 degrees of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.69

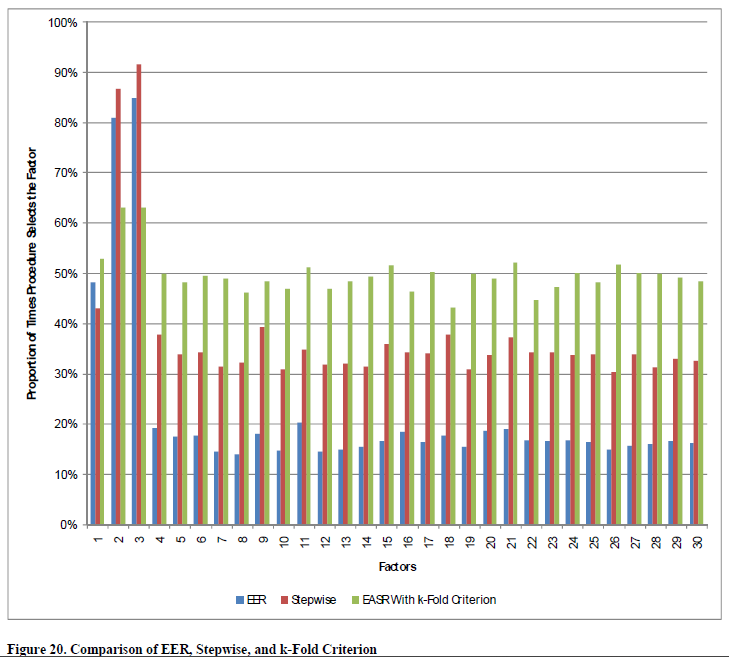
F-statistic: 196.7 on 6 and 513 DF, p-value: < 2.2e-16

A graph of blue squares

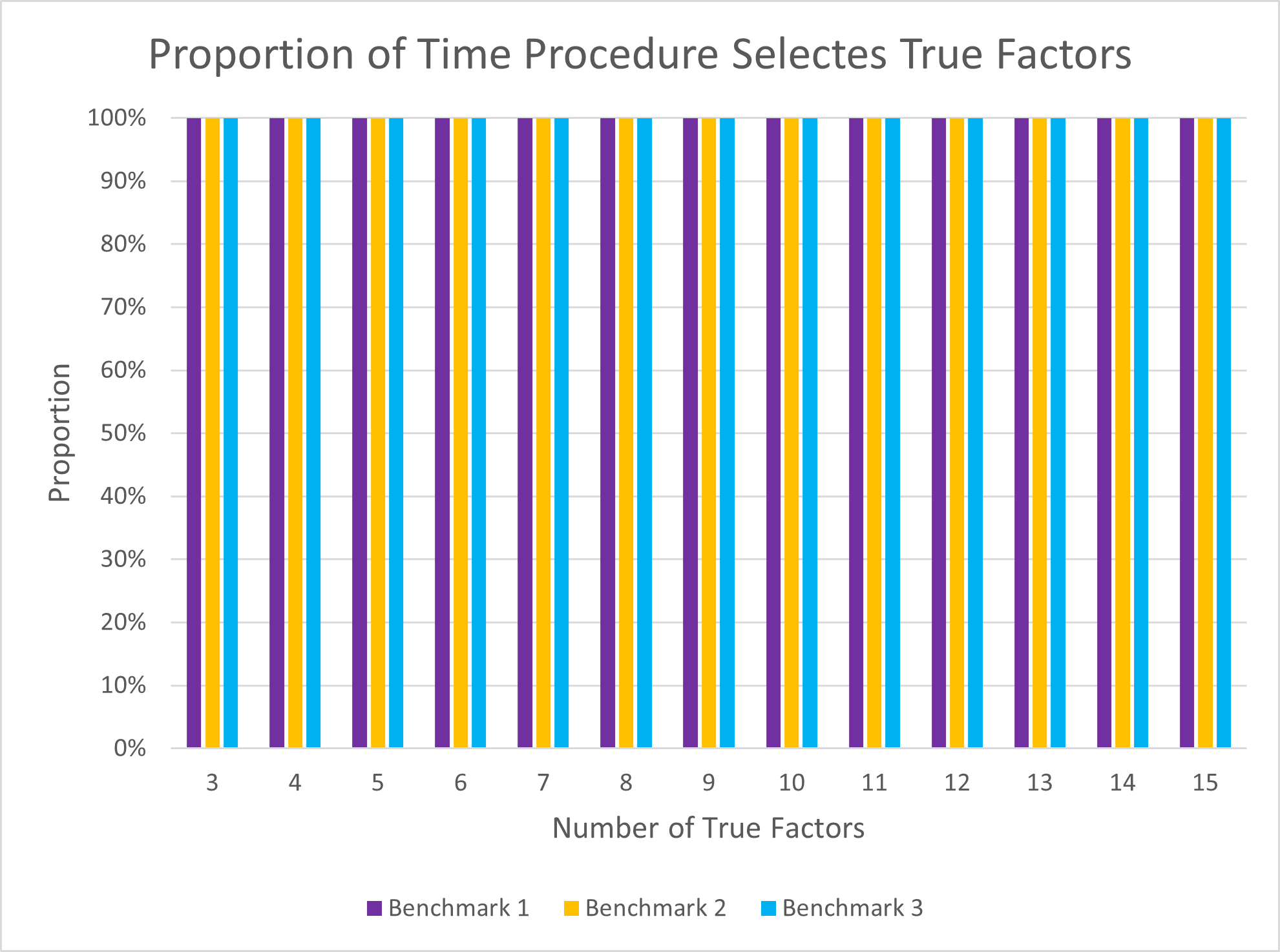
Description automatically generated with medium confidence

# Comparing the Performance of the Benchmarks

It is clear that our first benchmark variable selection algorithm, Lasso, performed the best out of the three given that it has the highest TNR across all 260k datasets of 0.867 (vs 0.800 for BM2 and 0.823 for BM2), it selected the highest number of correctly specified regression equations with 38,731 (vs 3,910 for BM2 and 4,822 for BM3), and it also has the lowest number of total (individual) False Positive factors, i.e. spurious regressors across all 260k datasets with 677,013 (vs 1,096,776 for BM2 and 942,252).

Original Figure 20

I re-created the equivalent of Figure 20 from the original Working Paper

My version of it for these results

The same chart using True Negative Rates instead so it gives more information

