Best Subset Selection: Some Recommendations for Practitioners

Maya Lozinski

November 2018

Contents

I. Introduction

In many regression problems, more variables are available than can or should be used. But which subset of variables should be included and which should be excluded? One version of this problem is the best subset selection problem, i.e. which 10 (or 20 or 100) variables should one choose from a large set of possible variables to maximize a model's explanatory power?

The widely used Lasso is a relaxation of the best subset selection problem. It produces a regression result where many coefficients are exactly zero, thus excluding the associated variables from the model. However, lasso also returns estimates biased towards zero, and, as a result, has undesirable properties in highly correlated datasets. Lasso is popular in part because, unlike best subset selection, it can be solved quickly in large, high dimensional datasets. However, recent work has made the best subset selection problem newly tractable in much larger datasets than before (Bertsimas et al. 2016).

However, many practical considerations remain unanswered in regards to best subset selection. In particular, how the solutions compare in terms enforcing a level of sparsity which reflects the true data generating process? What are the tradeoffs in choice of solution? In addition, can the lasso result be used to more efficiently choose a hyperparameter for the best subset selection problem?

II. Policy Relevance:

My own interest in these questions arises from my research on health policy. Best subset selection has several useful features from a policy perspective.

First, it is highly similar to ordinary least squares regression (in fact, unconstrained best subset selection *is* OLS regression). OLS regression models have been used to set payments and penalties in many health policy settings for the last 15+ years. As such, the best subset selection produces regression results that are substantially more interpretable to policy makers than lasso results.

In addition, lasso regression creates a political economy problem. As mentioned before, OLS regression is used to set payments in multiple healthcare settings. For example, Medicare uses an OLS regression model to adjust health insurance premium payments for the health conditions of beneficiaries. An insurance company might get paid an extra $5000 per year for covering a patient with diabetes. However, a lasso estimator would shrink the coefficient for diabetes, and thus potentially shrink the payment. As such, a lasso estimator (or any estimator with bias towards zero) is a political non-starter. Social scientists have skirted around this bias issue by running OLS, which is unbiased, on the variables selected by lasso regression. However, best subset selection can solve this problem more directly.

III. Background on Sparse Methods

Best Subset selection and Lasso are similar problems, but differ in a few key ways, outlined below:

Best Subset Selection

- Solution to: $\min_{B} \frac{1}{2} \|y - XB\|_2^2 \; s.t. \; \|B\|_0 \leq k$ (Count of nonzero B values is less than k)
- Sparsity: Number of nonzero variables can be chosen directly.
- Computation: NP Hard, but Bertsimas et al. 2016 provide Mixed Integer Optimization (MIO) formulations that are feasible for moderate size problems

Lasso

- Solution to: $\min_{B} \frac{1}{2} \|y - XB\|_2^2 \; s.t. \; \|B\|_1 \leq t$ (Sum of absolute values of B is less than t)
- Sparsity: Cannot directly choose number of nonzero variables; results from hyperparameter choice.
- Computation: Multiple algorithms, fast and feasible in settings with large *p* and *n*.

Previous work (Hastie et al. 2017) has compared the two sparse methods in terms of predictive performance. They find that neither best subset selection nor the lasso uniformly dominate the other in terms of prediction accuracy. Best subset selection generally performs better when the signal to noise ratio is high; lasso generally performs better when it is low.

IV. Data and Methods

For this analysis, I use multiple simulated datasets to compare the regressions. Simulated data can vary on infinite dimensions. Here I allow for random variation on the dimensions I consider to be most crucial for external validity – the true coefficients vary size and there is noise in the outcome. In one set of datasets, the variables are also arbitrarily correlated.

Data: Simulated data for two types of datasets, as described below.

- Base Data:
  - Data Generating Equation: $Y_i = BX_i + \epsilon_i$
  - k of p coefficients ($\beta \sim U(-1,1)$) are nonzero. True coefficients vary in size and sign.
  - $X_i,\ \epsilon_i \sim N(0,1)$
- Correlated Data: Covariates are arbitrarily correlated; all other features the same as the base data

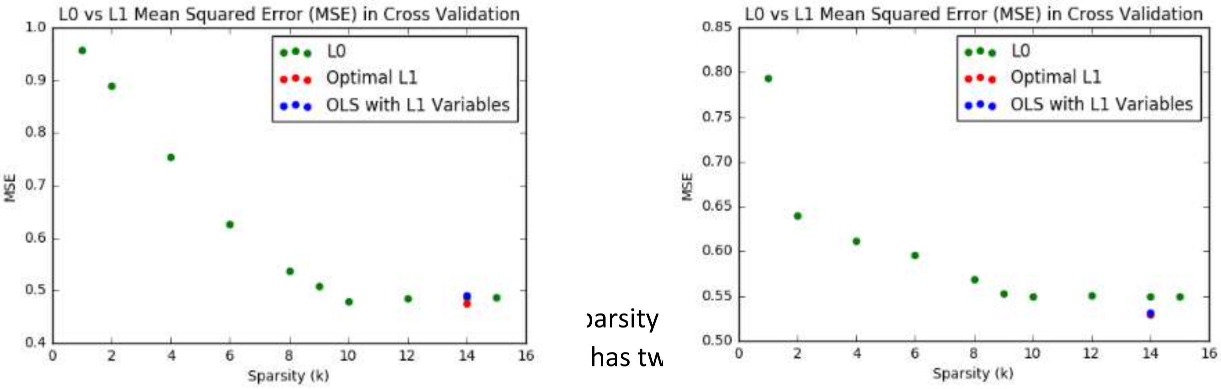
Implementation:

- Implemented Best Subset Selection regression problem with Gurobi Optimizer Version 8.0 as a Mixed Integer Programming problem, as formulated by Bertsimas et al. 2016. s
- Used Python 3.6.4 with standard packages, including Panda, Matplotlib, and scikitlearn to interface with Gurobi and perform analysis.

V. Results and Recommendations

i. **Recommendation 1:** Use an F-test or the one-standard deviation rule to select sparsity (k) in cross-validation.

Table 1: Cross validation mean squared error (MSE) in data with (i) independent and (ii) correlated covariates (k=10, p=15, n=1000). MSE is graphed for Best Subset Selection at a range of sparsity choices (k). Lasso and Post- Lasso cross validation error is graphed at the cross validation choice of λ.



(i) Base data

(ii) Correlated Data

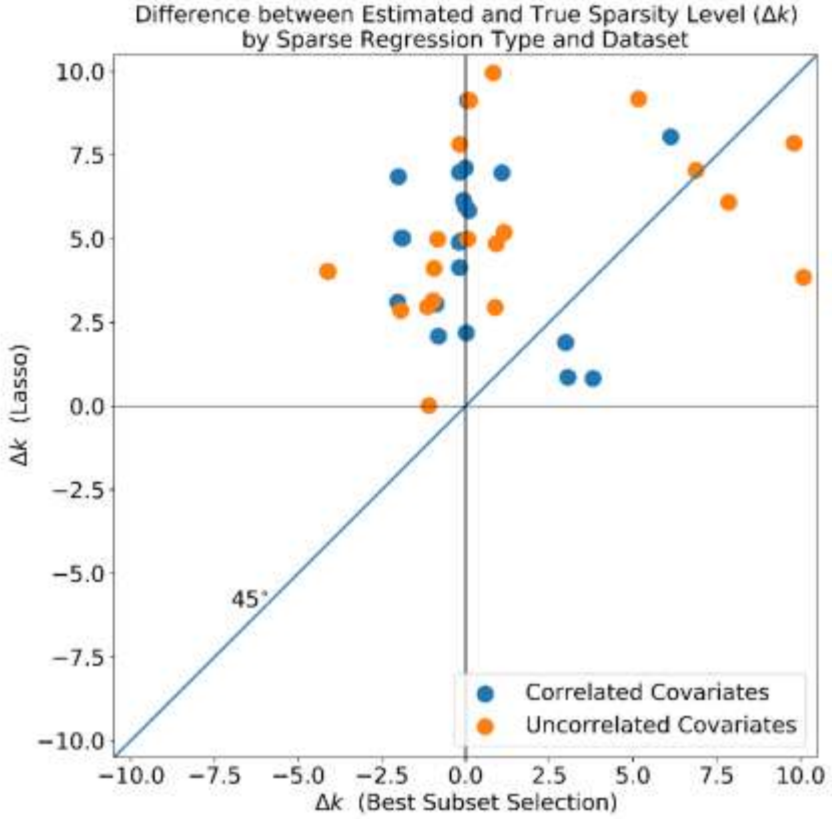
ɔarsity
has tw

First, it is better to choose a larger k when the true k is unknown if predictive performance is very important.

Second, cross validation has substantial scope to overestimate (but not underestimate) k*. One should consider using an F-test to determine if each increase in k significantly improves prediction. In contrast, traditional cross validation involves selecting the k with lowest MSE. Alternatively, use the one standard deviation rue to select k*.

ii. **Recommendation 2:** Lasso sparsity can provide a warm start to search over the space of sparsity choices (k).

Table 2: Difference between actual and estimated sparsity (k) for Lasso and Best Subset Selection from cross-validation. Cross-validation criteria was the minimum mean-squared error. Each point represents an independent simulation ((k=10, p=20, n=1000)). Orange dots represent the base data generating process and blue dots represent the correlated data generating process.

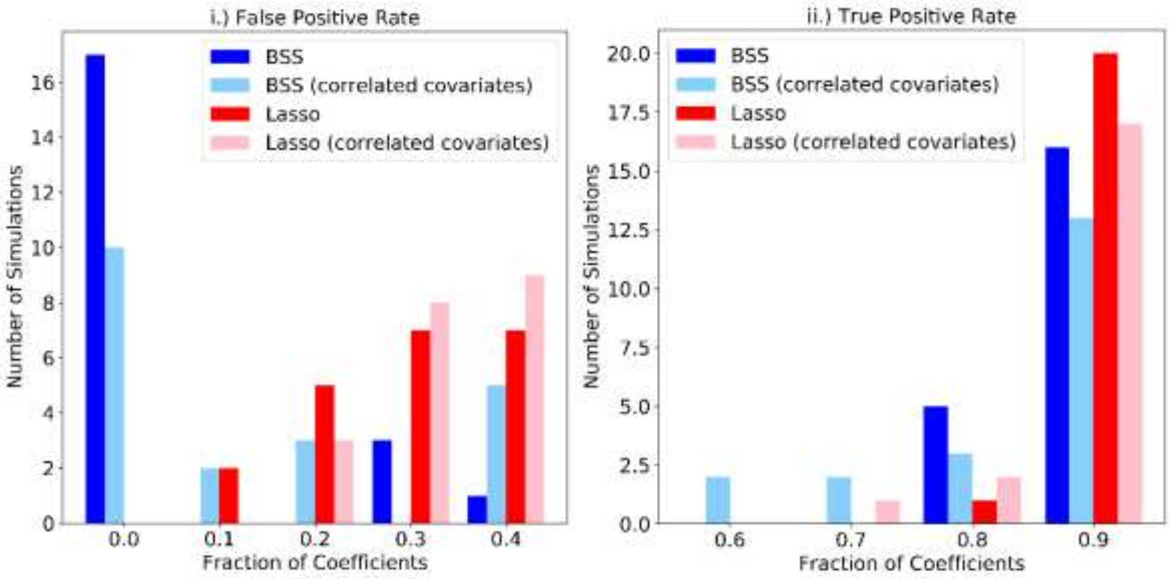


Discussion: Note that the Lasso sparsity tends to be larger than both the BSS k and the true k. The blue line represents the 45 degree line and points above the line have a larger lasso k than BSS k. Lasso nearly always overestimates the true k. In contrast, BSS estimates are relatively, but not entirely, centered on the true k.

Focus the initial search for BSS Sparsity to the left of the Lasso sparsity. For example, search for the optimal choice of k starting with the lasso k and incrementing down. Here [Note this will link to the attached document called "efficient-subset-selection"]is an algorithm for one possible fast search heuristic, along with sufficient conditions to achieve the correct results.

iii. **Recommendation 3:** Choose best subset selection over lasso in applications where false positives more costly than false negatives.

Table 3: For 20 simulations with the base and correlated data generating processes (k=10, p=20, n=1000), plotted below are i.) False Positive Rate ii.) True Positive Rate



Discussion: Best subset selection results in much less Type I error (false positives) at the cost of some Type II error (false negatives). BSS has a lower false positive rate than lasso, a desirable property. However, it also a lower true positive rate, so it is more likely to incorrectly set coefficients to zero. Neither regression type strictly dominates the other. The choice of which to use depends on the whether false positives or false negatives are worse in a given application.

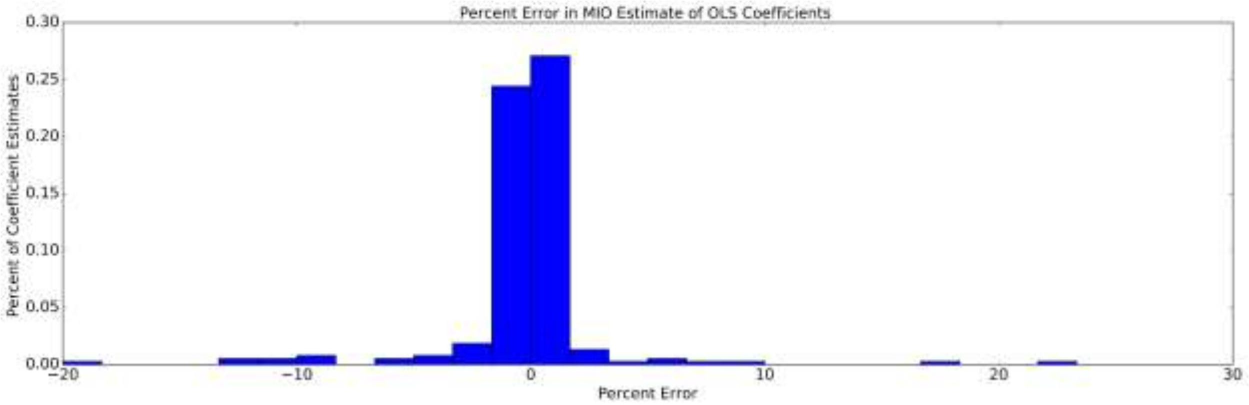
Applications where Best Subset Selection is preferable to Lasso:

- Data with substantial multicollinearity - Fewer nonzero coefficients will result in more stable coefficients (and therefore smaller standard errors) due to less collinearity.
- Weak Instruments – Some have proposed using Lasso in settings with many instrumental variables to weed out weak instruments. In this context, false positives are highly undesirable, as weak instruments will bias the results. BSS should perform better at excluding weak instruments from the regression than Lasso, leasing to less biased results.
- Hypothesis Generation – Regressions are sometimes used to generate hypotheses, which will them be investigated through other scientific means. For example, in genetics research, sparse methods can be used to determine which genes may be contribute to a specific biological function and guide future lab work. If investigating each hypothesis is time consuming, then the BSS should be used because it generates fewer false positives.

iv. Recommendation 4: Post-process best subset selection estimates using the closed form OLS formula.

Observation 4: As noted before, speedy best subset selection requires mixed integer optimization (MIO). While MIO solvers accurately choose the best subset of coefficients, they estimate those coefficients with some degree of error relative to closed form OLS estimates.

Table 4: Percent error in coefficient estimates from MIP solver (in comparison to closed form OLS estimates) conditional on same subset of nonzero variables.



Discussion: The error is close to zero for most estimates, making this inaccuracy trivial in many cases. However, a few coefficient estimates are off by a substantial amount.

Therefore, as a best practice, one should post-process best subset selection estimates. Recompute the nonzero coefficients using the closed form OLS formula (i.e. $\hat{B} = (X'X)^{-1}X'Y$).

VI. Limitations:

This work has several limitations. Most notably, this analysis uses only simulated data, although efforts were made to replicate features of real data. The data generating process was chosen to replicate many of dimensions heterogeneity found in real world data, including variable effect sizes and arbitrary correlation between variables. However, there are undoubtably exceptions for which this data is not representative. Future work could extend the analysis presented here to a greater variety of datasets. In addition, the future work should investigate in greater depth search procedures and heuristics for faster cross validation.

VII. Conclusion

Many best subset selection estimators are newly feasible, raising practical questions about their usage. This work attempts to address considerations about hyper-parameter choice, cross-validation, and relative performance of best subset selection. The results have several implications for practitioners using best subset selection methods. First, prediction quality can be quite sensitive to the choice of k, suggesting there are benefits to choosing more generous values of k. In addition, cross-validation appears to provide reasonable scope to overestimate k. However, Lasso may provide a strong "warm start" to the search for k_0 (the empirically optimal k), should one want to use best subset selection. Lastly, MIO solvers do not return the exact OLS coefficients, so practitioners should recompute them using the standard formula.

VIII. References

Bertsimas, Dimitris, Angela King, and Rahul Mazumder. "Best subset selection via a modern optimization lens." The annals of statistics 44, no. 2 (2016): 813-852.

Hastie, Trevor, Robert Tibshirani, and Ryan J. Tibshirani. "Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso." arXiv preprint arXiv:1707.08692 5 (2017). *Scikit learn*

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12, no. Oct (2011): 2825-2830.