Statistical Science
2020, Vol. 35, No. 4, 625–626
https://doi.org/10.1214/20-STS733REJ
Main article: https://doi.org/10.1214/19-STS701, https://doi.org/10.1214/19-STS733

Rejoinder: Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons

Trevor Hastie, Robert Tibshirani and Ryan J. Tibshirani

With two papers under discussion, our task in this rejoinder is somewhat unusual in that we have been asked not to comment directly on the paper by Bertsimas, Pauphilet and Van Parys (BPV). We would like to congratulate BPV, however, for making impressive computational advances in best subsets (BS) and related problems.

There is a great deal of interesting material in the discussions, and these provide excellent background and perspectives on the sparse regression problem. Our rejoinder will be brief and will touch on some common themes that were expressed by this distinguished group of scientists.

- Sarwar, Sauk and Sahinidis (SSS) give an impressive detailed summary of the setups and conclusions from both papers. George gives a nice history of how these two papers evolved and a detailed overview of the findings from both papers. He mentions promising Bayesian alternatives, for example, "spike and slab" methods. The big question is whether they can be made computationally scalable. For example, Johndrow, Orenstein and Bhattacharya (2020) present some exciting work in the direction of scalability.
- Generally, the discussants agree that the lasso tends to work better in low SNR settings, while BS excels with high SNR ones. Importantly, in our paper we go beyond the lasso, and study a simple form of the relaxed lasso, while in their paper, BPV generalize BS to an $\ell_0 + \ell_2$ penalty. Both generalizations result in methods that work well across a broader SNR spectrum; BPV's combined penalty also leads to a considerable speedup of the BS algorithm, while the relaxed lasso extension in our paper adds very little cost to the computation of the lasso.
- Mazumder argues that high SNR problems do exist in some areas of business, signal processing and image classification. We acknowledge this, but point out that in many high SNR settings (such as image classification), linear models would rarely be used, and nonlinear adaptive methods such as gradient boosting and neural networks would probably be preferred. He also makes the important point that when $p \gg n$, (e.g., $n = 20$ and $p = 1000$) and the SNR is 6, the *achieved* PVE (test set proportion of variance explained) by the lasso is only about 0.02. This is despite the fact that the population PVE is $6/(6 + 1) \approx 0.86$. This underlines the fact that SNR does not fully capture the difficulty of a supervised learning problem.
- Mazumder also discusses his own "regularized subsets" proposal (Hazimeh and Mazumder, 2018) which adds an ℓ_1 or ℓ_2 regularizer to the BS problem. This appears to be the same or similar to the BPV proposal, but seems to not have been mentioned by these other authors or other discussants.
- SSS comment on the $\ell_0 + \ell_2$ speedup in the BPV paper, pushing back against the claim by BPV that their resulting procedure is within one or two orders of magnitude as fast as the lasso via `glmnet`. SSS say that "this appears to be comparing the time needed to find a single cardinality-constrained solution set to the full solution path for `glmnet`". We agree: if you multiply the "1 or 2 orders of magnitude" by the 100 values along the path, the difference becomes more like 3 or 4 orders of magnitude.
- Chen, Taeb and Buhlmann (CTB) give a very nice survey of subset selection techniques, and enrich this series by introducing several other ways to judge the different methods. CTB wondered why we had not considered the adaptive lasso (Zou, 2006), along with the relaxed lasso, since it also pushes the solution toward a less-regularized alternative. This is a reasonable suggestion, but unlike the relaxed lasso, the adaptive lasso does not provide a full continuum of less-regularized estimates from the lasso to least squares. In the same vein, we did not include "SparseNet" (Mazumder, Friedman and Hastie, 2011) in the main comparisons in our paper (though it is included in the supplement) which is even more closely aligned with

Trevor Hastie is Professor, Departments of Statistics and Biomedical Data Science, Stanford University, Stanford, California 94305, USA (e-mail: hastie@stanford.edu). Robert Tibshirani is Professor, Departments of Statistics and Biomedical Data Science, Stanford University, Stanford, California 94305, USA (e-mail: tibs@stanford.edu). Ryan J. Tibshirani is Associate Professor, Departments of Statistics and Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania 15217, USA (e-mail: ryantibs@cmu.edu).

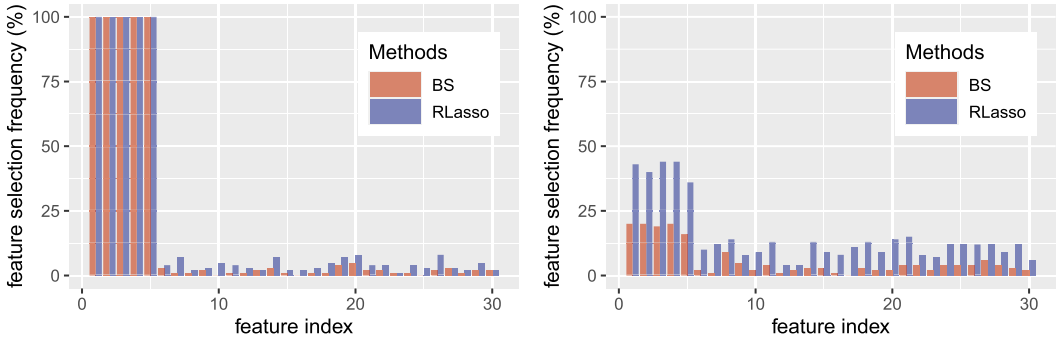

FIG. 1. Figure 3 of Chen, Taeb and Buhlmann, with the relaxed lasso used in place of the lasso.

methods in between ℓ_1 and ℓ_0. As we mention in our Introduction, there is a long list of potential competitors. We chose just four methods: ℓ_0 and ℓ_1, forward stepwise (FS) which has a long history as an easy alternative to ℓ_0, and the relaxed lasso for its ease of implementation.

- We strongly agree with SSS about the importance of software to provide functionality for different approaches "under one hood." Since the initial submission of our paper, we released version 4.0 of `glmnet`, adding functionality for the relaxed lasso. It includes a `relax=TRUE` option in the call, which triggers the additional computation of the relaxed fits, and `cv.glmnet` will then choose both λ and γ. The software also allows for any GLM family via the alternative `family=family()` specification, which extends this methodology to a much broader scope of problems.

  Using this new package, we reran CTB's Figure 3 on feature stability, using the relaxed lasso in place of the usual lasso. We see in Figure 1 that the relaxed lasso significantly reduces the number of false positives compared to the lasso, while maintaining the superiority over best subsets in selecting the true variables in the low SNR regime.
- Figure 3 of SSS's discussion seems to show that the relaxed lasso slows down considerably beyond about $n = 1500$ in their setup. However, we would like to point out that this speed comparison is misleading, as different methods consider different model sizes—this being just a function of the range of tuning parameters used in this example. In particular, in the dense part of its solution path, the relaxed lasso considers models of size approaching n (in this example, 2000, while SSS limited forward-selection to a maximum of 100 steps (variables). Thus, while both methods are computing least squares fits, they are done on very different numbers of variables, and thus not comparable.
- Finally, we note that some of the discussants had trouble exactly reproducing some of the BPV results exactly. It was for this reason that we provided a public GitHub repo https://github.com/ryantibs/best-subset, with fully reproducible R code for all of our examples. We encourage all authors to do this in the future.

We would like to thank the Editors of *Statistical Science* and the discussants for their considerable efforts. Hopefully, these papers and discussions represent, in total, a useful contribution toward a topic that is a central one in statistical practice.

REFERENCES

HAZIMEH, H. and MAZUMDER, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper. Res.* **68** 1517–1537.

JOHNDROW, J. E., ORENSTEIN, P. and BHATTACHARYA, A. (2020). Bayes shrinkage at GWAS scale: Convergence and approximation theory of a scalable MCMC algorithm for the horseshoe prior. *J. Mach. Learn. Res.*. To appear.

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769 https://doi.org/10.1198/jasa.2011.tm09738