Statistical Science
2020, Vol. 35, No. 4, 609–613
https://doi.org/10.1214/20-STS808
Main article: https://doi.org/10.1214/19-STS701, https://doi.org/10.1214/19-STS733

Modern Variable Selection in Action: Comment on the Papers by HTT and BPV

Edward I. George

Let me begin by congratulating the authors of these two papers, hereafter HTT and BPV, for their superb contributions to the comparisons of methods for variable selection problems in high dimensional regression. The methods considered are truly some of today's leading contenders for coping with the size and complexity of big data problems of so much current importance. Not surprisingly, there is no clear winner here because the terrain of comparisons is so vast and complex, and no single method can dominate across all situations. The considered setups vary greatly in terms of the number of observations n, the number of predictors p, the number and relative sizes of the underlying nonzero regression coefficients, predictor correlation structures and signal-to-noise ratios (SNRs). And even these only scratch the surface of the infinite possibilities. Further, there is the additional issue as to which performance measure is most important. Is the goal of an analysis exact variable selection or prediction or both? And what about computational speed and scalability? All these considerations would naturally depend on the practical application at hand.

The methods compared by HTT and BPV have been unleashed by extraordinary developments in computational speed, and so it is tempting to distinguish them primarily by their novel implementation algorithms. In particular, the recent integer optimization related algorithms for variable selection differ in fundamental ways from the now widely adopted coordinate ascent algorithms for the lasso related methods. Undoubtedly, the impressive improvements in computational speed unleashed by these algorithms are critical for the feasibility of practical applications. However, the more fundamental story behind the performance differences has to do with the differences between the criteria that their algorithms are seeking to optimize. In an important sense, they are being guided by different solutions to the general variable selection problem.

Focusing first on the paper of HTT, its main thrust appears to have been kindled by the computational breakthrough of Bertsimas, King and Mazumder (2016) (hereafter BKM), which had proposed a mixed integer optimization approach (MIO) for best subsets selection in problems with p as large as in the thousands. Requiring the optimization of ℓ_0-constrained least squares, conventional wisdom had long considered best subsets to be the computationally elusive gold standard for variable selection, having defied computation for p much larger than 30. Finally breaking this seemingly impenetrable barrier, MIO had suddenly unleashed a feasible implementation of best subsets for application in sparse high dimensional regression.

Illustrating the performance of MIO, BKM carried out simulation comparisons with some of its most prominent alternatives, including forward stepwise selection and the lasso. A close cousin of best subsets, stepwise is one of the most routinely used computable heuristic approximations for large p. The lasso, on the other hand, differs fundamentally from best subsets by its very nature. Obtained by optimizing an ℓ_1-penalized least squares criterion rather the best subsets ℓ_0-constrained criterion, it substitutes a rapidly computable convex optimization problem for an NP-hard nonconvex optimization problem. The BKM simulations demonstrated setups where best subsets substantially dominated both stepwise and the lasso in terms of both predictive squared error loss and variable selection precision, appearing to confirm the gold standard promise of best subsets.

Concerned that BKM's simulation terrain was too limited to come to such a universal conclusion, HTT set out to perform broader simulation comparisons. In particular, the terrain of comparisons has been expanded to include setups with a broader range of SNRs. As opposed to BKM, HTT now include setups with weaker SNRs corresponding to PVE values that more realistically characterize applications often encountered in practice. In addition to comparing best subsets, stepwise and the lasso, HTT include a new contender, the (simplified) relaxed lasso, driven by an interesting combination of the lasso and least squares.

From this broader terrain of comparisons presented by HTT, new patterns of relative performance emerge. To begin with, the performance of stepwise is now appears very similar to that of best subsets throughout. The major differences between stepwise and best subsets found by BKM disappear when stepwise is tuned by cross-validation (here on external validation data) rather than AIC. This is valuable to see because the choice of stopping rule has been controversial for applications of step-

Edward I. George is the Universal Furniture Professor of Statistics and Economics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA (e-mail: edgeorge@wharton.upenn.edu).