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).

wise in practice. HTT's comparisons suggest that cross-validated stopping may be the best way to go.

More importantly, HTT's performance evaluations highlight interesting differences between best subsets and both the lasso and relaxed lasso. In predictive performance, the lasso now appears better than best subsets for smaller SNRs, but worse for larger SNRs with a performance crossover that varies from setup to setup. The same is true for HTT's selection accuracy metrics, though the lasso regularly produces too many nonzero estimates especially in the higher SNR settings. An exception where best subsets dominates throughout occurs when the known mutual incoherence (MI) conditions for lasso consistency seem to fail. Remarkably, the relaxed lasso fares extremely well throughout the comparisons. It appears to adaptively move closer to the best of the methods for each SNR value, occasionally even outperforming them all. Although no one of the four procedures is best in every setting, the relaxed lasso certainly comes the closest. Finally, it is notable that in terms of explained variation, namely PVE, no essential differences between the four methods appeared.

The observed performance differences between these methods can be intuitively understood as stemming from how their induced coefficient estimators are driven to balance bias-variance tradeoffs in order to minimize prediction error on the validation data. For best subsets and stepwise, this induced estimator is simply least squares on a selected subset of coefficients. For the lasso, it is a shrinkage adjusted least squares estimator which has been soft-thresholded by a selected λ. And for the relaxed lasso, it is a γ-weighted mixture of a λ-soft-thresholded lasso estimator with a least squares estimator on the lasso's nonzero coordinates, for a selected λ and γ. Note that each of these induced estimators consists of both zero and nonzero estimates of the components of the actual underlying β. Thus, the squared prediction error over the validation set will include the squared bias of the zero estimates in addition to the variance and squared bias of the nonzero estimates.

From this bias-variance tradeoff perspective, it is straightforward to see why best subsets and stepwise provide desirable variable selection and prediction when the variance of the subset least squares estimates is small enough relative to the sizes of the underlying nonzero coefficients. This is exactly what occurs throughout HTT's setups as SNR increases (with β and n fixed) and the subset least squares estimates become more and more accurate. However, when the variance of subset least squares is relatively large compared to the sizes of the nonzero coefficients, which is what occurs in HTT's setups where SNR is small, the accuracy and stability of the least squares estimates can be substantially improved by shrinkage. This is precisely where we see the lasso improvements over best subsets and stepwise. However, there is also a notable conflict between the goals of variable selection and prediction for the lasso. With the choice of λ guided by prediction error, the lasso is forced to shrink less in order not to over-bias the nonzero estimates of the larger sized coefficients. This is what is leading to the lasso's increased number of nonsparse estimates that appears in so many of the plots.

It is very interesting to understand how the relaxed lasso is able to better negotiate the tension between selection and prediction faced by the lasso. Combining the lasso and least squares by unshrinking the nonzero lasso estimates back toward the least squares estimates, what makes the relaxed lasso so effective is that it does this adaptively. At first glance, it may seem puzzling that the relaxed lasso would yield fewer nonzero estimates than the lasso. However, this is explained by the fact that both λ, the amount of lasso shrinkage, and γ, its mixing parameter, are being simultaneously tuned by prediction error on external validation data. By controlling the shrinkage of its nonzero estimates for better predictions, the relaxed lasso can balance increasing thresholding of the smaller estimates with less biasing of the larger estimates. Via the external validation data, it adapts this balance to the data at hand. Thus the the relaxed lasso is able to offer good accuracy and prediction error across all the SNR settings.

In a nutshell, the least squares component of the relaxed lasso is serving to stabilize the estimates of the larger size coefficients so that they do not interfere with the lasso variable selection, thereby resulting in fewer false nonzero coefficient estimates. It may be of interest to note that this aspect of the relaxed lasso is similar in spirit to the scalable, fully Bayesian spike-and slab lasso (SSL) of Ročková and George (2018). The SSL also stabilizes the larger size estimates, but by using the log of a mixture of spike-and-slab Laplace priors as a regularization criterion. In effect, the SSL adaptively applies simultaneous strong thresholding lasso shrinkage to smaller coefficients and weak lasso shrinkage to stabilize the larger coefficients. By including a prior on the mixing weights, the SSL becomes even more self adaptive, avoiding the need to use cross-validation for mixing weight estimation.

Turning now to focus on the paper by BPV, which followed the paper by HTT, its main thrust revolves around CIS and SS, two newer approaches introduced by the authors in Bertsimas, Pauphilet and Van Parys (2017) and Bertsimas and Van Parys (2020). In contrast to the ℓ_0-constrained best subsets criterion of MIO, both CIS and SS are driven by an ℓ_0-constrained, ℓ_2-penalized least squares criterion for variable selection in linear regression. CIS does this with a cutting-plane algorithm for fast integer optimization, while SS applies a dual subgradient algorithm to a continuous Boolean relaxation of the criterion. These algorithms are impressive and effective advances, especially in terms of their speed and ability to handle larger problems. As seen across the simulations,

CIS and SS offer dramatic improvements in speed over MIO, coping now with problems of size $n = 1000$ and $p = 20{,}000$ within seconds of the glmnet implementation of the lasso. Apparently, they can even scale further to feasibly handle problems of sizes n, p of 100,000 or $n = 10{,}000$ and $p = 1{,}000{,}000$ within minutes, a major step forward.

Beyond this increased feasibility for larger problems, which is obviously a huge boon for practicality, the statistical value of CIS and SS ultimately rests on the inferential consequences of adding an ℓ_2-penalization to the ℓ_0-constrained least squares criterion of MIO. With this addition, an ℓ_2-penalized subset least squares estimator, adaptively tuned by cross-validation, is now induced to reduce prediction error. Such ℓ_2-penalization was implicit in the early shrinkage proposals of ridge regression and Stein estimation for improved prediction as well as stabilization of least squares estimation (Hoerl and Kennard, 1970, Stein, 1960, James and Stein, 1961). But in contrast to ℓ_1 induced shrinkage, ℓ_2-penalization alone does not lead to thresholding and so is not directly useful for selection. However, when combined with the ℓ_0-constrained criterion as BPV have done, the improved accuracy of the induced coefficient estimator can be expected to lead to both improved variable selection and prediction, especially in settings where the variance of least squares is large relative to the sizes of the nonzero regression coefficients. It may be of interest to note that such thresholding of an ℓ_2 shrinkage estimator is akin to the automatic thresholding by the positive-part James–Stein estimator, which also has the attractive theoretical property of being classically minimax under squared predictive loss. Further, more powerful minimax versions of such positive-part shrinkage estimators for selection were developed by Zhou and Hwang (2005).

Illustrating the statistical potential of CIS and SS, BPV go on to compares their performance to three prominent shrinkage-selection methods that use enhanced lasso related penalties, namely ENet (the elastic net), MCP (maximum convex penalty) and SCAD (smoothly clipped absolute deviation). ENet enhances the lasso criterion with the addition of an ℓ_2-penalization, in the same way that CIS and SS enhance the best subsets criterion. More precisely, ENet is driven by a mixture of ℓ_1 and ℓ_2 penalties, parametrized by a shrinkage parameter λ and a mixing parameter α, both of which are then tuned by cross-validation. The combination of the ℓ_1 and ℓ_2 penalization serves to stabilize the shrinkage of subset least squares, especially in the context of highly correctly predictors. On the other hand, both SCAD and MCP are driven by carefully tailored nonconvex relaxations of the ℓ_1-criterion that automatically allow for lighter shrinkage of the estimates of the larger coefficients, in that way mitigating the tension between selection and prediction faced by the lasso. Both MCP and SCAD are parametrized by both a shrinkage parameter λ and a shape parameter γ which are adaptively tuned to the data via cross-validation. Just as for the HTT comparisons, the statistical performance differences between all these methods can be intuitively understood from the bias-variance tradeoffs faced by their induced coefficient estimators as they are predictively tuned by cross-validation.

BPV carry out simulated comparisons of CIS, SS, ENet, MCP and SCAD across various versions of the basic underlying Gaussian structure used by HTT. For their terrain of comparisons, they consider six noise/correlation settings with three noise levels: low (SNR $= 6$, $k = 100$, $p = 20{,}000$), medium (SNR $= 1$, $k = 50$, $p = 10{,}000$) and high (SNR $= 0.05$, $k = 10$, $p = 2000$), and two predictor correlation levels: $\rho = 0.2, 0.7$. Although there are lots of moving parts across these setups, one can say roughly that the low noise settings, where the signal is easiest to measure, are the least challenging for variable selection, whereas the high noise settings, where the signal is most hidden, are the most challenging and perhaps most like what is most often encountered in practice. As elaborated by HTT, both the low and medium noise SNR values correspond to PVE values that may be less frequently in encountered in practice. It should also be noted that unlike HTT, BPV vary the sample size n within each of their settings, thereby also illustrating performance differences as n increases. Revealing the limiting behavior of the methods, increasing n has the effect of making the problems statistically easier in terms of measuring the signal. By decreasing the implicit variance of the subset least squares estimates on which shrinkage is applied, increasing n thus diminishes the importance of shrinkage for each the methods. In this way, the effect of increasing n parallels the effect of increasing SNR.

Beyond these settings, BPV also provide valuable comparisons on simulated data where the MI conditions for lasso consistency fail, as well as comparisons using data generated with a real design matrix X with $p = 14{,}858$ gene expressions for $n = 1145$ lung cancer patients. With this X, they simulate hypothetical continuous patient outcomes from various Gaussian linear models with SNR levels varying from 0.05 to 6, as well as discrete 0–1 patient outcomes by thresholding these linear models. Extending the methods for the classification data, they apply CIS and SS driven by hinge loss, and ENet, MCP and SCAD driven by logistic loss.

For performance evaluations of the competing methods, BPV highlight their potential for variable selection with accuracy A (the fraction of true nonzero found), FDR, (the false discovery rate), and ROC curves of true positives versus false positives. Taken together these reveal the selection tradeoffs obtained by each of the methods. For predictive performance, they report MSE, casting prediction

primarily as a "practical implication of accuracy." And to convey the practicality of their CIS and SS implementations, they also report speed comparisons across various settings, gauging the speed of every method against the gold standard fastest speed of the lasso.

Across all these simulation comparisons, many different patterns of relative accuracy, false discovery rates and prediction quality emerge. Beginning with the accuracy and MSE comparisons when the actual underlying sparsity level k_{true} is treated as fixed and known, so that $A = 1 - \text{FDR}$. This constraint forces all the methods to explicitly trade correct selection for false detections. It is hence not surprising that ENet is the least accurate as it also suffers from the lasso's tension between selection and prediction. In contrast it is impressive that both CIS and MCP are the most accurate throughout, MCP apparently avoiding this tradeoff tension with its relaxed penalty. More pronounced in the low noise and high correlation settings, these differences diminish rapidly as we would expect when n and/or the noise level is increased, virtually disappearing in the high noise setting. For prediction in the low and medium noise settings, CIS is best, slightly beating SS and MCP and SCAD, and with ENet clearly worst, though again all these differences seem to disappear in the more statistically difficult high noise setting. The speed comparisons here demonstrate that CIS and SS are dramatically faster than MIO, and even computationally comparable with MCP and SCAD for larger n. Indeed, in the high noise setting, SS appears to be nearly as fast as ENet, which is as fast as the lasso in every setting.

For the more realistic settings where the sparsity level k_{true} is unknown and estimated via cross validation, we first see from the ROC curves, that CIS, SS and MCP essentially provide the most desirable tradeoffs between true and false selection in the low and medium noise settings, with ENet performing worst. However, these differences become much less pronounced and even reversed in the high noise setting, especially when the predictor correlations are high. For prediction in the low and medium noise settings, CIS continues to be best up to around k_{true} with MCP becoming best at the higher levels of estimated sparsity, with ENet worst throughout. It is interesting that the predictions of MCP, SCAD and ENet continue to improve with increasing k, suggesting that they add more variables to improve their predictive cross-validation tuning. In the high noise setting, it appears that MCP and SCAD predict best, with ENet second under high correlation, although the wide error bars for these comparisons suggest such differences may not be real. In terms of accuracy, ENet, MCP and SCAD are best, followed by CIS and then SS in the low and medium noise settings, all differences disappearing as n increases. In the high noise setting, these differences are more pronounced and persistent. However, in terms of FDR, CIS is clearly best and ENet persistently worst in all settings. SS and MCP are also sometimes best, improving along with SCAD as n increases. As BPV suggest, this overall observed tendency of ENet, MCP and SCAD to select many more variables than are selected by CIS and SS, can be understood to be a consequence of using soft penalization rather than a hard cardinality constraint to drive variable selection and prediction.

For the settings where the MI conditions for lasso consistency fail, Enet is seen, not surprisingly, to be among the least accurate across all three noise settings with an accuracy converging to less than 1 for large n. In contrast, the accuracies of the other four methods, which do not rely on the MI conditions, converge to 1 in the low and medium noise settings. CIS and SS are best in all settings, along with MCP which is also one of the best in the low noise setting. CIS and SS are also seen to dominate in prediction in the low and medium noise settings. However, in the high noise setting, the predictive performances of all the methods become indistinguishable.

Lastly, we consider the simulated comparisons based on the real data design matrix where the predictors correlations are apt to be less symmetrically structured and perhaps more realistic. For the linear model setup, as SNR increases, ENet is increasingly the least accurate with the most false detections, while MCP is increasingly the most accurate with the fewest false detections. However, these differences all disappear at the lowest SNR levels where the variable selection problem is most challenging. In terms of the ROC curve tradeoffs, ENet goes from best to worst and MCP goes from worst to best as the SNR level is increased from 0.05 to 6. For the classification model setup, we see ENet, MCP and SCAD generally providing better accuracy compared to CIS and SS across all the noise settings, with the differences diminishing as n, the noise level and the predictor correlations are increased. But in terms of false detection rates, this ordering is reversed, except for MCP which also obtains lower false detection in the low noise settings. For the classifications obtained by their respective cross-validated support selections, SS appears to provide the most desirable tradeoffs between accuracy, false detection and sparsity.

In concluding, let me thank the HTT and BPV for their illuminating comparisons of some of the leading approaches for variable selection in high dimensional regression, and for their many insightful summary conclusions with which I agree. Together, they have provided convincing evidence that the relaxed lasso, CIS and SS will be very valuable additions to the arsenal of modern variable selection approaches. Of course, as so aptly demonstrated throughout the simulations, no single approach will dominate all others in every setting, especially as pertains to the ease of detecting the underlying signal which is typically unknown. So in practice it makes

good sense to consider a collection of methods together, for example, beginning an application with lasso predictive screening and then proceeding from there with other methods, as BPV have suggested. Finally, it should be emphasized that the arsenal of valuable modern variable selection approaches goes well beyond the methods considered here, as HTT have indicated. In particular, in line with the underlying Bayesian nature of penalized likelihood methods, a growing number of promising scalable Bayesian variable selection approaches, for example, the spike-and-slab lasso described earlier, are valuable additions to the arsenal and should not be overlooked.

ACKNOWLEDGMENT

This work was supported by NSF Grant DMS-1916245.

REFERENCES

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

BERTSIMAS, D., PAUPHILET, J. and VAN PARYS, B. (2017). Sparse Classification: a scalable discrete optimization perspective. Under revision.

BERTSIMAS, D. and VAN PARYS, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Ann. Statist.* **48** 300–323. MR4065163 https://doi.org/10.1214/18-AOS1804

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 https://doi.org/10.1080/01621459.2016.1260469

STEIN, C. M. (1960). *Multiple Regression. Essays in Honor of Harold Hotelling*. Stanford Univ. Press, Palo Alto, CA.

ZHOU, H. H. and HWANG, J. T. G. (2005). Minimax estimation with thresholding and its application to wavelet analysis. *Ann. Statist.* **33** 101–125. MR2157797 https://doi.org/10.1214/009053604000000977