Statistical Science
2020, Vol. 35, No. 4, 623–624
https://doi.org/10.1214/20-STS701REJ
Main article: https://doi.org/10.1214/19-STS701, https://doi.org/10.1214/19-STS733

Rejoinder: Sparse Regression: Scalable Algorithms and Empirical Performance

Dimitris Bertsimas, Jean Pauphilet and Bart Van Parys

We would like to thank the discussants of our paper [6] for their insightful comments and the authors in [11] for their excellent work that collectively enhance our understanding of sparse regression. We would like to use this rejoiner as an opportunity to emphasize some observations made by these authors and that contribute, in our opinion, to the dialog and provide some new insights. Throughout, we use the notation in [6].

1. Sarwar et al. [16] provide a thorough and unified comparison of the methods presented in both [6] and [11], with an emphasis on computational tractability and software. We are glad to see mixed-integer optimization (MIO) approaches well represented in their benchmark analysis. Yet, a decade ago, MIO techniques would not have been included. Indeed, the sparse linear regression problem

$$\min_{w\in\mathbb{R}^p} \|y - Xw\|^2 + \frac{1}{2\gamma}\|w\|_2^2 \quad \text{s.t. } \|w\|_0 \le k, \tag{1}$$

is $\mathcal{NP}$-hard. Unfortunately, the theory of $\mathcal{NP}$-hardness established in the early 1970s has contributed to the belief in many scientific communities that discrete optimization problems were intractable, which at the time and until the mid-1990s was by and large justified. Since the early 1990s, however, the field of MIO has made significant advances in our ability to model and solve high dimensional problems [4]. As far as best subset selection is concerned, the papers [4, 6, 7] establish that approaching sparsity exactly via MIO is computationally feasible for $n \sim 200{,}000$ and $p \sim 100{,}000$. Sarwar et al. [16] also included recent work [13, 14] that further extend the scalability to $n, p \sim 10^6$. These significant computational developments force us, in our opinion, to rethink the beliefs established in 1970 s that discrete optimization problems are intractable. On the contrary, we believe that many cardinality problems in the field could now benefit from advances in MIO [2] and that discrete optimization methods should be included in the curriculum of graduate programs in statistics and machine learning.

2. In their excellent discussion, Chen et al. [9] provide a clarifying and structured survey of feature selection methods. In their "additional thoughts," they emphasize alternative objectives to assess the quality of the selected features, such as distributional robustness and stability. As they recall, [1, 18] show that Lasso is equivalent to a robust linear regression problem

$$\begin{aligned}\min_{w} \max_{\Delta\in\mathcal{U}(\lambda)} &\|y - (X+\Delta)w\|_2 \\ &= \min_{w}\big(\|y - Xw\|_2 + \lambda\|w\|_1\big),\end{aligned} \tag{2}$$

with $\mathcal{U}(\lambda) = \{\Delta : \|\Delta_i\|_2 \le \lambda\}$, where Δ_i is the ith column of the matrix Δ. Specifically, Lasso assumes that the columns of X are subject to adversarial noise Δ_i that is restricted to satisfy $\|\Delta_i\|_2 \le \lambda$, $i \in [p]$ and finds coefficients w that minimize the worst-case error $\|y - (X+\Delta)w\|_2$. In other words, Lasso attempts to immunize (robustify) linear regression against perturbations in the data. Note that similar results exist for ridge regression but with a different noise model, that is, with $\mathcal{U}(\lambda) = \{\Delta : (\sum_{i\in[n],j\in[p]} \Delta_{ij}^2)^{1/2} \le \lambda\}$. Moreover, unlike theorems about the feature selection ability of Lasso [8, 15, 17, 19], equation (2) universally holds without any assumptions on the data. Together, this collection of results clearly indicates that ℓ_1-regularization undeniably provides robustness and sometimes (depending on stringent assumptions that are hard to verify) accurately selects features. However, in the statistics community at large, ℓ_1-regularization is mostly advertised and used as a method that primarily induces sparsity (see [12] for an overview). This belief plays down the importance of robustness in the practical success and relevance of Lasso. Lasso is able to withstand noisy data in a very mathematically precise way, and real-world data is noisy. Moreover, viewing Lasso as a feature selection term only suggests excluding from the ℓ_1 penalty covariates that are known to be part of the support (e.g., the intercept). By doing so, the corresponding coefficients are no longer immunized against noise. As rightfully observed by Chen et al. [9], the search for sparsity should not outshine the need for robustness.

Dimitris Bertsimas is the Boeing Professor of Operations Research, Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue Bldg. E62-560, Cambridge, Belmont, Massachusetts 02139, USA (e-mail: dbertsim@mit.edu). Jean Pauphilet is a Doctoral Student, Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue Bldg. E40, Cambridge, Belmont, Massachusetts 02139, USA (e-mail: jpauph@mit.edu). Bart Van Parys is an Assistant Professor, Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue Bldg. E62-569, Cambridge, Belmont, Massachusetts 02139, USA (e-mail: vanparys@mit.edu).

3. George [10] provides solid intuition on how the relaxed Lasso from [11] and the $\ell_0 - \ell_2$ estimators considered in [6] both balance feature selection with prediction error. In the case of the $\ell_0 - \ell_2$ estimators (1), this tradeoff is explicitly modeled via the ℓ_0 constraint $\|w\|_0 \leq k$ (for sparsity) and the ℓ_2 penalty $\frac{1}{2\gamma}\|w\|_2$ (for robustness). While Lasso addresses robustness only, through ℓ_1-regularization, the relaxed Lasso proposed by Hastie et al. [11] appears as an effective and tractable alternative which addresses the deficiencies of Lasso regarding feature selection. Beyond the intuition, it would be interesting to investigate how these two properties, sparsity and robustness, come into play in the relaxed Lasso formulation and theoretically evaluate its performance. Besides sparsity and robustness, other desirable properties are sought after in statistical estimators, such as group sparsity, limited pairwise multicollinearlity, automatic detection of nonlinear transformations, and statistical significance. On this regard, MIO modeling could be used to simultaneously take into account these desirable properties, and whenever it is not possible to satisfy all these properties simultaneously, provide a guarantee that it is indeed impossible to do so. We refer to [2], Chapter 5 and [3, 5] for the development of this holistic regression framework.

CONCLUSIONS

In summary, we would like to thank the editorial team, the anonymous reviewers and the discussants for their contributions. We hope this issue will provide new insights and guidance on feature selection in statistics, as well as emphasize the importance that mixed-integer and robust optimization should play in the future of the field.

REFERENCES

[1] BERTSIMAS, D. and COPENHAVER, M. S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European J. Oper. Res.* **270** 931–942. MR3814540 https://doi.org/10.1016/j.ejor.2017.03.051

[2] BERTSIMAS, D. and DUNN, J. (2019). *Machine Learning Under a Modern Optimization Lens*. Dynamic Ideas Press, Belmont, MA.

[3] BERTSIMAS, D. and KING, A. (2016). OR forum—an algorithmic approach to linear regression. *Oper. Res.* **64** 2–16. MR3463258 https://doi.org/10.1287/opre.2015.1436

[4] 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

[5] BERTSIMAS, D. and LI, M. L. (2020). Scalable holistic linear regression. *Oper. Res. Lett.* **48** 203–208. MR4077406 https://doi.org/10.1016/j.orl.2020.02.008

[6] BERTSIMAS, D., PAUPHILLET, J. and VAN PARYS, B. (2020). Sparse regression: Scalable algorithms and empirical performance. *Statist. Sci.* **35** 555–578.

[7] 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

[8] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846 https://doi.org/10.1002/cpa.20124

[9] CHEN, Y., TAEB, A. and BÜHLMANN, P. (2020). A Look at Robustness and Stability of ℓ_1-versus ℓ_0-Regularization: Discussion of Papers by Bertsimas et al. and Hastie et al. *Statist. Sci.* **35** 614–622.

[10] GEORGE, E. (2020). Modern variable selection in action: Comment on the papers by HTT and BPV. *Statist. Sci.* **35** 609–613.

[11] HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. (2020). Best subset, forward stepwise, or lasso? Analysis and recommendations based on extensive comparisons. *Statist. Sci.* **35** 579–592.

[12] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

[13] HAZIMEH, H. and MAZUMDER, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Available at arXiv:1803.01454.

[14] HAZIMEH, H., MAZUMDER, R. and SAAB, A. (2020). Sparse regression at scale: Branch-and-bound rooted in first-order optimization. Available at arXiv:2004:06152.

[15] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[16] SARWAR, O., BENJAMIN SAUK, B. and SAHINIDIS, N. (2020). A discussion on practical considerations with sparse regression methodologies. *Statist. Sci.* **35** 593–601.

[17] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

[18] XU, H., CARAMANIS, C. and MANNOR, S. (2009). Robustness and regularization of support vector machines. *J. Mach. Learn. Res.* **10** 1485–1510. MR2534869

[19] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449