

# 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 rejoinder 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

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

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

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

with  $\mathcal{U}(\lambda) = \{\Delta : \|\Delta_i\|_2 \leq \lambda\}$ , where  $\Delta_i$  is the  $i$ th column of the matrix  $\Delta$ . Specifically, Lasso assumes that the columns of  $X$  are subject to adversarial noise  $\Delta_i$  that is restricted to satisfy  $\|\Delta_i\|_2 \leq \lambda$ ,  $i \in [p]$  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  $\mathcal{U}(\lambda) = \{\Delta : (\sum_{i \in [n], j \in [p]} \Delta_{ij}^2)^{1/2} \leq \lambda\}$ . 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  $\ell_1$ -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,  $\ell_1$ -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  $\ell_1$  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](mailto: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](mailto: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](mailto:vanparys@mit.edu)).*