Projected-L0 Decoder for Variable Selection in Linear Regression

Maxim Fedotov, David Rossell, Gábor Lugosi

Universitat Pompeu Fabra

June 2024

There are plenty of methods known to be useful for variable selection in a linear regression setup. In particular, there has been a battle between ℓ_0 and ℓ_1 methods in the literature, and a couple of bright representatives of these approaches are the best subset selection and the Lasso.

The best subset problem is known to be NP-hard, see Natarajan (1995), whereas the Lasso solves a convex optimization problem, and thus should be easier to compute in general. Even though Hastie et al. (2017) claim that neither method dominates the other with best subset selection performing generally better in high signal-to-noise ratio regimes and the Lasso performing better otherwise, the Lasso is known to fail in selecting the active subset of variables with probability at least $1/2$ if a so-called mutual incoherence condition does not hold, see Wainwright (2009b). More importantly, there exist nearly isotropic design matrices which violate mutual incoherence, see Wainwright (2009a), and thus the Lasso does not recover the active set with non-negligible probability in these cases.

So, the interest for ℓ_0 methods is natural even though they may be computationally demanding. In particular, Wainwright (2009a) analyses an "optimal decoder" over all subsets of covariates of size s – the size of the active set – and shows that it can succeed in terms of variable selection even when the Lasso fails. It becomes even more exciting in the light of the result from Bertsimas et al. (2016) which shows a way to state the best subset problem as a mixed integer optimization problem for which there exist solvers like Gurobi. This advancement shrinks the boundaries of the best subset problem instances which were thought to be practically unsolvable.

We introduce a two step Projected-L0 method for variable selection which is supposed to be less computationally demanding than the optimal decoder from Wainwright (2009a) which we refer to as Exact-L0. At the first stage of the algorithm, a winner is chosen among subsets of the same size according to a projected ℓ_0 criterion, thus leaving at most $\min\{n, d\}$ subsets for consideration, where n is the number of observations and d is the total number of variables.

At the second stage, an exact ℓ_0 criterion with a penalty is used on the preselected subsets. The projected ℓ_0 criterion used to select the best model for each sensible size is just a mean squared error taken at a sub-vector of a precomputed global estimator. The performance of the algorithm depends on the properties of the global estimator. For example, one can run the Lasso over all the covariates available, and use it as the global estimator, which makes sense when Lasso does not concentrate on the active set right away. In this case, the Projected-L0 can refine the result obtained via the Lasso.

Considering the subsets of size s, we show that Projected-L0 with the Lasso global estimator consistently estimates the active set of covariates even when the Lasso fails, i.e. when mutual incoherence does not hold. We show that the type 1 error probability of the Projected-L0 decays exponentially in n for all $n \geq n_0$, where n_0 is a sufficient sample size. We also compare the error probability to the one of Exact-L0 and identify the loss occurring due to approximation. Moreover, we show how each tournament among subsets of the same size at the pre-selection stage can be formulated as a binary quadratic problem with a single linear constraint.

References

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813 – 852, 2016. doi: 10.1214/15-AOS1388. URL `https://doi.org/10.1214/15-AOS1388`.

Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009a.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_0-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009b.