In Defense of l_0

Dongyu Lin DONGYU@WHARTON.UPENN.EDU
Emily Pitler EPITLER@SEAS.UPENN.EDU
Dean P. Foster DEAN@FOSTER.NET
Lyle H. Ungar UNGAR@CIS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104 USA

Keywords: variable selection, best subset selection, l_1 regularization, Lasso, stepwise regression

Abstract

In the past decade, there has been an explosion of interest in using l_1-regularization in replace of l_0-regularization for feature selection. We present results showing that while l_1-regularization never outperforms l_0-regularization by more than a constant factor, in some cases using an l_1 penalty is infinitely worse than an l_0 penalty. We also compare algorithms solving these two problems from several aspects and show that, although good solutions have been developed for l_1 problem, they may not perform as well as the very classic stepwise regression, which is a greedy l_0 surrogate. In other words, "an approximate solution to the right problem" can be better than "the exact solutions to the wrong problem".

We focus on variable selection problems in which there is a large set of potential features, only a few of which are likely to be helpful. This type of sparsity occurs often in various machine learning tasks, such as predicting disease based on millions of genes, or predicting the topic of a document based on the occurances of hundreds of thousands of words.

Consider a normal linear model

$$\mathbf{y} = X\beta + \varepsilon,$$

where $\mathbf{y}$ is the response variable with n observation, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)'$ is the coefficient parameters, and error $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. Assume that only a subset of $\{\mathbf{x}_j\}_{j=1}^p$ has nonzero coefficients.

The traditional statistical approach to this problem, namely, the l_0 problem, finds an estimator that minimizes the l_0 penalized sum of squared errors

$$\arg\min_\beta \left\{ \|\mathbf{y} - X\beta\|^2 + \Pi\sigma^2 \|\beta\|_{l_0} \right\}, \tag{1}$$

where $\|\beta\|_{l_0} = \sum_{i=1}^p I_{\{\beta_i \neq 0\}}$.

However, this problem is shown to be NP hard (Natarajan, 1995). An trackable problem relaxes the l_0 penalty by the l_1 norm $\|\beta\|_{l_1} = \sum_{i=1}^p |\beta_i|$ and seeks

$$\arg\min_\beta \left\{ \|\mathbf{y} - X\beta\|^2 + \lambda \|\beta\|_{l_1} \right\}, \tag{2}$$

known as the l_1-regularization problem (Tibshirani, 1996). The computation of (2) is much more efficient and approachable because of the convexity (Efron et al., 2004; Candes & Tao, 2007).

There are three main reasons that we think l_0 problem is the more correct problem and that stepwise regression can perform better than l_1 algorithms in sparse cases.

First, l_0 solutions are more predictive.

Suppose $\hat{\beta}$ is an estimator of β. We consider the predictive risk of $\hat{\beta}$, which is defined as

$$R(\beta, \hat{\beta}) = E_\beta \|X\beta - X\hat{\beta}\|^2. \tag{3}$$

We consider the special case when X is orthogonal. The l_0 problem can be solved by simply picking those predictors with least squares estimate $|\hat{\beta}_i| > \gamma \geq 0$, where the choice of γ depends on the noise level of the model and the penalty scale Π in (1). (Donoho & Johnstone, 1994; Foster & George, 1994) showed that $\Pi = 2\log p$ is optimal.

Let

$$\hat{\beta}_{l_0}(\gamma_0) = (\hat{\beta}_1 I_{\{\hat{\beta}_1 > \gamma_0\}}, \ldots, \hat{\beta}_p I_{\{\hat{\beta}_p > \gamma_0\}})' \tag{4}$$

be the l_0 estimator that solves (1). and the l_1 solution to (2)

$$\hat{\beta}_{l_1}(\gamma_1) = \Big(\text{sign}(\hat{\beta}_1)(|\hat{\beta}_1| - \gamma_1)_+, \ldots,$$

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

$$\text{sign}(\hat{\beta}_p)(|\hat{\beta}_p| - \gamma_1)_+\Big)', \tag{5}$$

where $\hat{\beta}_i$'s are the least squares estimates.

We have the following theorems:

Theorem 1. *For any $\gamma_1 \geq 0$, there exists constants $C_1 > 0$ and $C_2 > 0$, such that*

$$\inf_{\gamma_0} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_0}(\gamma_0))}{R(\beta, \hat{\beta}_{l_1}(\gamma_1))} \leq 1 + C_1 \cdot \gamma_1 e^{-C_2 \gamma_1}. \tag{6}$$

This theorem implies that the worse l_0 solution performs almost as good as the best l_1 solution in both saturated ($\lambda_1 \approx 0$) and sparse ($\lambda_1 \gg 0$) cases.

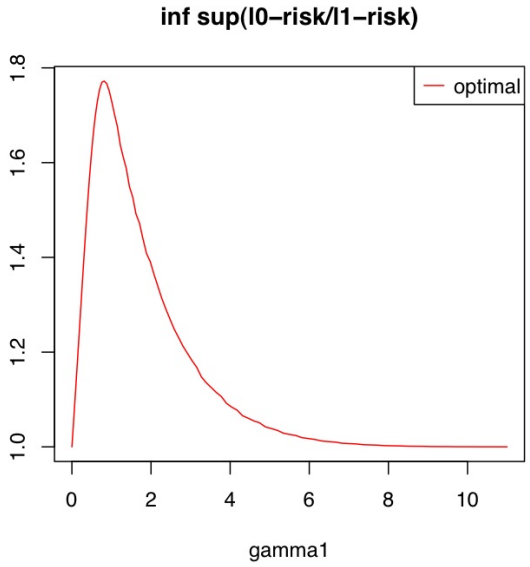

Figure 1. $\inf_{\gamma_0} \sup_{\beta} \frac{R(\beta,\hat{\beta}(\gamma_0))}{R(\beta,\hat{\beta}(\gamma_1))}$ tends to 1 when $\gamma_1 \to 0$ or ∞. Specifically, the supremum over γ_1 is bounded.

On the contrary,

Theorem 2. *For any γ_0, there exists constants $C_3 > 0$ and $r > 0$, such that*

$$\inf_{\gamma_1} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_1}(\gamma_1))}{R(\beta, \hat{\beta}_{l_0}(\gamma_0))} \geq 1 + C_3 \cdot \gamma_0^r. \tag{7}$$

It diverges when $\gamma_0 \to \infty$, in other words, when the system is extremely sparse, l_1 solution will do a terrible job.

Recall that $\gamma_0 = \sqrt{2 \log p}$ is optimal, in which case, the above minimax ratio will blow up when p is very large.

The reason for this is because $\hat{\beta}_{l_1}$ is a biased estimate of β and is shrunk towards zero a lot when the system is sparse, as shown in Figure 3.

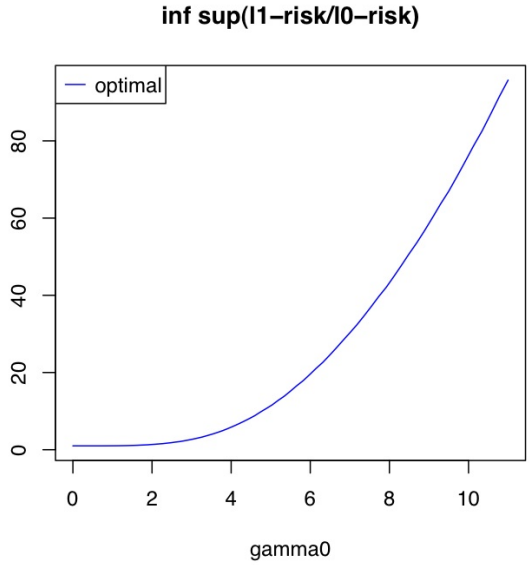

Figure 2. $\inf_{\gamma_1} \sup_{\beta} \frac{R(\beta,\hat{\beta}(\gamma_1))}{R(\beta,\hat{\beta}(\gamma_0))}$ tends to ∞ when $\gamma_0 \to \infty$.

These results show that l_0 solution provides a less risky and more predictive solution. Empirical results also show that l_0 is substantially better than l_1 if there are detectable signals (George & Foster, 2000; Foster & Stine, 2004; Zhou et al., 2006)

Second, l_0 controls FDR better.

The False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) is defined as $E[V/R|R > 0]P(R > 0)$ where R is the total number of discoveries and V is the number of false discoveries among them.

(Abramovich et al., 2006) shows that an FDR-penalized procedure is adaptively optimal in any l_p ball, $0 \leq p \leq 2$. For small $\|\beta\|_{l_0}$, this penalty has the flavor of an l_0 penalty. Also, the solution is indeed a variable hard threshold rule. Hence in some sense, when a sparse solution is preferred, hard thresholding, or l_0 solution surpasses other solutions.

We also compare the empirical FDR $\hat{V}/\hat{R}$ and forward stepwise regression does a better job in controlling FDR than Lasso does.

Third, the l_0-based stepwise regression provides sparser solutions.

Compared to l_0-regularization, l_1 does not always provide the sparsest possible solution. It is easy to construct an example where l_1 will pick a solution that with smaller l_1 norm but a lot more nonzero coefficients (Candes et al., 2007).

However, the NP-hardness makes l_0 problem un-

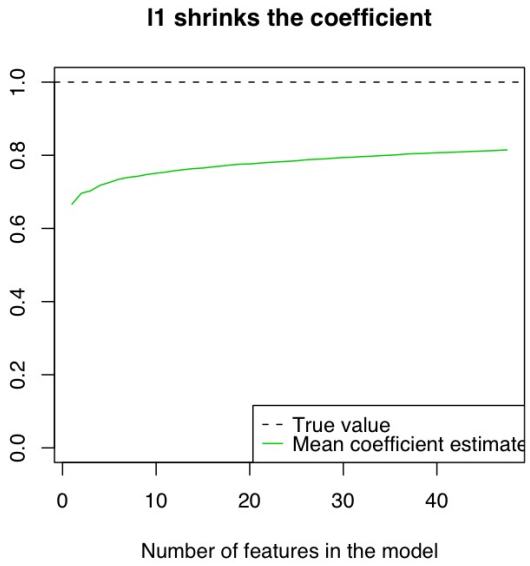

Figure 3. The true β is 1 while $\hat{\beta}_{l_1}$ is always shrunk by at least 20%.

tractable. (Natarajan, 1995) reduced the known NP hard problem of "the exact cover of 3-sets" to the best subset selection problem. It is then fair to ask which comes closer to solving this type of problem: a greedy approximation to the l_0 problem or an exact solution to the l_1 problem? It turns out that forward stepwise regression gets not only sparser but also more accurate results than Lasso does.

Conclusion

An approximation to the correct problem is better than the exact solutions to a wrong problem.

References

Abramovich, F., Benjamini, Y., Donoho, D. L., & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, *34*, 584–653.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, *57*, 289–300.

Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, *35*.

Candes, E. J., Wakin, M. B., & Boyd, S. P. (2007). Enhancing sparsity by reweighted l_1 minimization.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*, 425–455.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.*, *32*.

Foster, D. P., & George, E. I. (1994). The Risk Inflation Criterion for Multiple Regression. *Ann. Statist.*, *22*, 1947–1975.

Foster, D. P., & Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *J. Amer. Statistical Assoc.*, *99*, 303–313.

George, E. I., & Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, *87*, 731–747.

Natarajan, B. (1995). Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, *24*, 227.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, *58*, 267–288.

Zhou, J., Foster, D. P., Stine, R. A., & Ungar, L. H. (2006). Streamwise feature selection. *J. Machine Learning Research*, *7*, 1861–1885.