A risk ratio comparison of L_0 and L_1 penalized regression

Dongyu Lin DONGYU@WHARTON.UPENN.EDU
Dean P. Foster DEAN@FOSTER.NET
Department of Statistics
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104, USA

Lyle H. Ungar UNGAR@CIS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor:

Abstract

In the past decade, there has been an explosion of interest in using l_1-regularization in place of l_0-regularization for feature selection. We present theoretical results showing that while l_1-penalized linear regression never outperforms l_0-regularization by more than a constant factor, in some cases using an l_1 penalty is infinitely worse than using an l_0 penalty. We also compare algorithms for solving these two problems and show that although solutions can be found efficiently for the l_1 problem, the "optimal" l_1 solutions are often inferior to l_0 solutions found using greedy classic stepwise regression. Furthermore, we show that solutions obtained by solving the convex l_1 problem can be improved by selecting the best of the l_1 models (for different regularization penalties) by using an l_0 criterion.

Keywords: Variable Selection, Regularization, Stepwise Regression

add citation to our Workshop paper Lin et al. (2008)

1 Introduction

In the past decade, a rich literature has been developed using l_1-regularization for linear regression including Lasso (Tibshirani, 1996), LARS (Efron et al., 2004), fused lasso (Tibshirani et al., 2005), grouped lasso (Yuan and Lin, 2006), relaxed lasso (Meinshausen, 2007), and elastic net (Zou and Hastie, 2005). These methods, like the l_0-penalized regression methods which preceded them Akaike (1973); Schwarz (1978); Foster and George (1994), address 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 is common in machine learning tasks, such as predicting disease based on thousands of genes, or predicting the topic of a document based on the occurrences of hundreds of thousands of words.

l_1-regularization is popular because, unlike the l_0 regularization historically used for feature selection in regression problems, the l_1 penalty gives rise to a convex problem that can be solved efficiently using convex optimization methods. l_1 methods have given reasonable results on a number of data sets, but there has been no careful analysis of how they perform when compared to l_0 methods. This paper provides a formal analysis of the two methods,

and shows that l_1 can give arbitrarily worse models. We offer some intuition as to why this is the case – l_1 shrinks coefficients too much and does not zero out enough of them – and suggest how to use an l_0 penalty with l_1 optimization.

We consider the classic normal linear model

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

with n observations $\mathbf{y} = (y_1, \ldots, y_n)'$ and p features $\mathbf{x}_1, \ldots, \mathbf{x}_p$, $p \gg n$, where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ "design matrix" of features, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ 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 regularization problem, finds an estimator that minimizes the l_0 penalized sum of squared errors

$$\arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda_0 \|\boldsymbol{\beta}\|_{l_0} \right\}, \tag{1}$$

where $\|\boldsymbol{\beta}\|_{l_0} = \sum_{i=1}^p I_{\{\beta_i \neq 0\}}$ counts the number of nonzero coefficients. However, this problem is NP hard Natarajan (1995). A tractable problem relaxes the l_0 penalty to the l_1 norm $\|\boldsymbol{\beta}\|_{l_1} = \sum_{i=1}^p |\beta_i|$ and seeks

$$\arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_{l_1} \right\}, \tag{2}$$

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

We assess our models using the predictive risk function (3)

$$R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \mathbb{E}_{\boldsymbol{\beta}} \|\hat{\mathbf{y}} - \mathbb{E}(\mathbf{y}|X)\|_2^2 = \mathbb{E}_{\boldsymbol{\beta}} \|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|_2^2. \tag{3}$$

We are interested in the ratios of the risks of the estimates provided by these two criteria. Unlike risk functions, predictive risk measures the true prediction error with irreducible variance from which noise has been removed. Smaller risks imply better expected performance in the future for prediction purpose. Recent literature has a focus on the selection consistency, where whether or the true variable can be identified is critical. However, in real application, due to the prevalent multicollinearity, highly correlated predictors are hard to separate from "true" or "false". Here we focus on the purpose of prediction accuracy and provocate [FORGET how to spell] the concept of predictive risk. **explain risk vs. consistency and the relation of risk to out-of-sample error; what else is this called b other people? see www-stat.wharton.upenn.edu/ stine/research/select.predRisk.pdf; maybe also Barbieri and Berger (2004)**

Our first result in this paper, given below as Theorems 1 and 2, is that l_0 estimates provide more accurate predictions than l_1 estimates do, in the sense of minimax risk ratios, as illustrated in Figure 1:

- $\inf_{\gamma_0} \sup_{\boldsymbol{\beta}} \frac{R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{l_0}(\gamma_0))}{R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{l_1}(\gamma_1))}$ is bounded by a small constant; furthermore, it is close to one for most γ_1s, especially for large γ_1s, which are mostly used in sparse systems.

- $\inf_{\gamma_1} \sup_{\boldsymbol{\beta}} \frac{R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}_{l_1}(\gamma_1))}{R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}_{l_0}(\gamma_0))}$ tends to infinity quadratically; in an extremely sparse system, the l_1 estimate may perform arbitrarily badly.
- $R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}_{l_1}(\gamma_1))$ is more likely to have a larger risk than $R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}_{l_0}(\gamma_0))$ does.

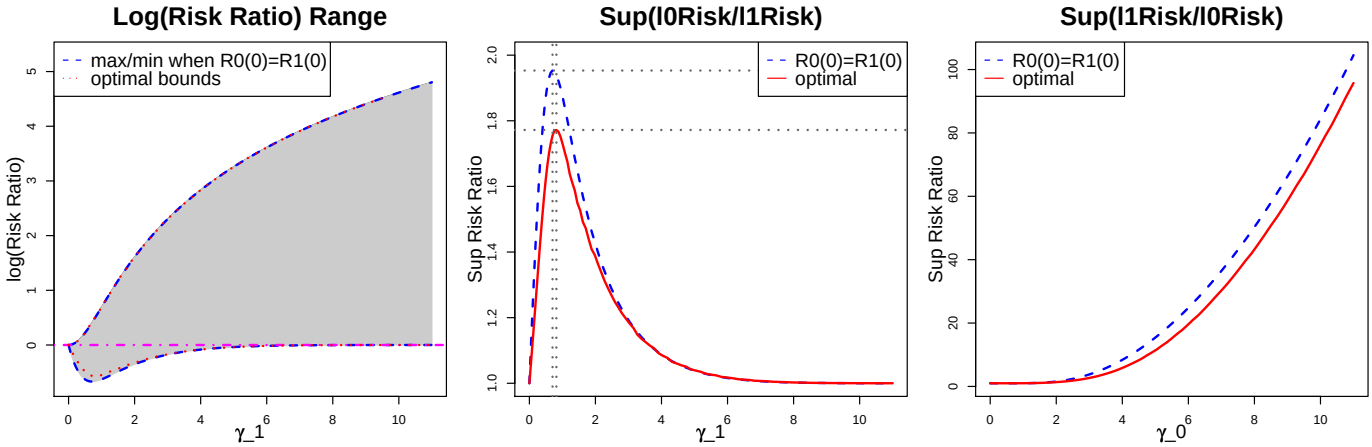

Figure 1: **Left**: The gray area shows the feasible region for the risk ratios–the log risk-ratio is above zero when l_0 produces a better fit. The graph shows that most of the time l_0 is better. The actual estimators being compared are those that have the same risk at $\beta = 0$, i.e., $R(0,\hat{\beta}_{l_0}(\gamma_0)) = R(0,\hat{\beta}_{l_1}(\gamma_1))$. **Middle**: This graph traces out the bottom envelope of the left hand graph (but takes the reciprocal risk ratio and no longer uses the logarithm scale). The dashed blue line displays $\sup_\beta R(\beta,\hat{\beta}_{l_0}(\gamma_0))/R(\beta,\hat{\beta}_{l_1}(\gamma_1))$ for γ_0 calibrated to have the same risk at zero as γ_1. This maximum ratio tends to 1 when $\gamma_1 \to 0$ (the saturated case) or ∞ (the sparse case). With an optimal choice of γ_0, $\inf_{\gamma_0} \sup_\beta R(\beta,\hat{\beta}_{l_0}(\gamma_0))/R(\beta,\hat{\beta}_{l_1}(\gamma_1))$ (solid red line) behaves similarly. Specifically, the supremum over γ_1 is bounded by 1.8. **Right**: This graph traces out the upper envelopes of the left hand graph on a normal scale. When $\gamma_0 \to \infty$, $\sup_\beta R(\beta,\hat{\beta}_{l_1}(\gamma_1))/R(\beta,\hat{\beta}_{l_0}(\gamma_0))$ tends to ∞, for both γ_1 that is calibrated at $\beta = 0$ and that minimizes the maximum risk ratio.

A detailed discussion on the risk ratios will be presented in Section 3, along with a discussion of other advantages of l_0 regularization. Our other results in the paper include showing that applying the l_0 criterion on an l_1 subset searching path can find the best performing model (Section 4) and running stepwise regression and Lasso on a reduced NP hard example shows that stepwise regression gives better solutions (Section 5).

We compare l_0 vs. l_1 penalties under three assumptions about the structure of the feature matrix X: independence, incoherence (near independence) and when the l_0 problem is NP-hard. For independence, we find: ... For near independence, we find that l_1 penalized regression followed by l_0 (explain) beats l_1, and for the NP-hard case, we find that if one could do the search, then the risk ratio could be arbitrarily bad for l_1 relative to l_0

2 Background on Risk Ratio

what is it, why is it good, where has it been used? risk vs consistence; and relation to out-of-sample error

3 Risk Ratio Results

3.1 l_0 solutions give more accurate predictions.

Suppose that $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Remember that the predictive risk of $\hat{\boldsymbol{\beta}}$ is defined as

$$R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \mathbb{E}_\beta \|\hat{\mathbf{y}} - \mathbb{E}(\mathbf{y}|X)\|_2^2 = \mathbb{E}_\beta \|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|_2^2.$$

We furthermore consider the case when X is orthogonal in this section. (For example, wavelets, Fourier transforms, and PCA all are orthogonal). The l_0 problem (1) can then be solved by simply picking those predictors with least squares estimates $|\hat{\beta}_i| > \gamma$, where the choice of γ depends on the penalty λ_0 in (1). It was shown Donoho and Johnstone (1994); Foster and George (1994) that $\lambda_0 = 2\sigma^2 \log p$ is optimal in the sense that it asymptotically minimizes the maximum predictive risk inflation due to selection.

Let

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

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

$$\hat{\beta}_{l_1}(\gamma_1) = \left(\text{sign}(\hat{\beta}_1)(|\hat{\beta}_1| - \gamma_1)_+, \ldots, \text{sign}(\hat{\beta}_p)(|\hat{\beta}_p| - \gamma_1)_+\right)', \tag{5}$$

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

We are interested in the ratios of the risks of these two estimates,

$$\frac{R(\beta, \hat{\beta}_{l_0}(\gamma_0))}{R(\beta, \hat{\beta}_{l_1}(\gamma_1))} \quad \text{and} \quad \frac{R(\beta, \hat{\beta}_{l_1}(\gamma_1))}{R(\beta, \hat{\beta}_{l_0}(\gamma_0))}$$

. I.e., we want to know how the risk is inflated when another criterion is used. The smaller the risk ratio, the less risky (and hence better) the numerator estimate is, compared to the denominator estimate. Specifically, a risk ratio less than one implies that the top estimate is better than the bottom estimate.

Formally, we have the following theorems, whose proofs are given in the last section:

Theorem 1 *There exists a constant C_1 such that for any $\gamma_0 \geq 0$,*

$$\inf_{\gamma_1} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_1}(\gamma_1))}{R(\beta, \hat{\beta}_{l_0}(\gamma_0))} \geq C_1 + \gamma_0. \tag{6}$$

I.e., given γ_0, for any γ_1, there exist β's such that the ratio becomes extremely large.

Contrast this with the protection provided by l_0:

Theorem 2 *There exists a constant $C_2 > 0$ such that for any $\gamma_1 \geq 0$,*

$$\inf_{\gamma_0} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_0}(\gamma_0))}{R(\beta, \hat{\beta}_{l_1}(\gamma_1))} \leq 1 + C_2 \gamma_1^{-1}. \tag{7}$$

The above theorems can definitely be strengthened, as demonstrated by the bounds shown in Figure 1, but at the cost of complicating the proofs. We conjecture that there exist constants $r > 1$, and $C_3, C_4, C_5 > 0$, such that

$$\inf_{\gamma_1} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_1}(\gamma_1))}{R(\beta, \hat{\beta}_{l_0}(\gamma_0))} \geq 1 + C_3 \gamma_0^r, \tag{8}$$

$$\inf_{\gamma_0} \sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_0}(\gamma_0))}{R(\beta, \hat{\beta}_{l_1}(\gamma_1))} \leq 1 + C_4 \gamma_1 e^{-C_5 \gamma_1}. \tag{9}$$

These theorems suggest that for any γ_1 chosen by the algorithm, we can always adapt γ_0 such that $\hat{\beta}_{l_0}(\gamma_0)$ outperforms $\hat{\beta}_{l_1}(\gamma_1)$ most of the time and loses out a little for some β's; but for any γ_0 chosen, no γ_1 can perform consistently reasonably well on all β's.

Because of the additivity of risk functions, (see appendix equations (14) and (15)), due to the orthogonality assumption, we focus on the individual behavior of β for each single feature. Also the risk functions are symmetric on β, so only the cases of $\beta \geq 0$ will be displayed.

Figure 2 illustrates that given γ_1, we can pick a γ_0, s.t. the risk ratio is below 1 for most β except around $(\gamma_0 + \gamma_1)/2$, yet this ratio does not exceed one by more than a small factor, even for the worst case.

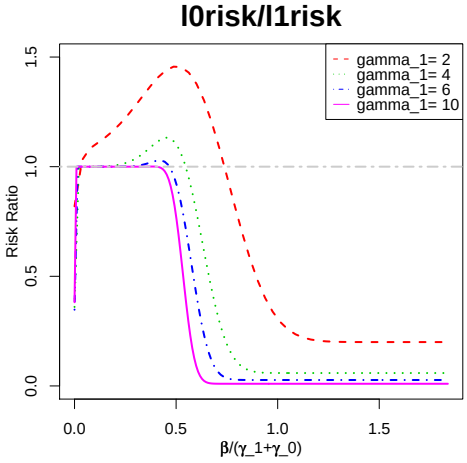

Figure 2: For each γ_1, we let $\gamma_0 = \gamma_1 + 4\log(\gamma_1)/\gamma_1$. This choice of γ_0 makes the risk ratios small at $\beta \approx 0$ and $\beta \geq \gamma_0$, only inflated around $\beta/(\gamma_0 + \gamma_1) = 1/2$, albeit very little especially when γ_1 is large enough.

The intuition as to why l_0 fares better than l_1 in the risk ratio results is that l_1 must make a "devil's choice" between shrinking the coefficients too much or putting in too many spurious features. l_0 penalized regression avoids this problem. This section explains this in more detail.

3.2 l_1 shrinks coefficients too much

Why does the l_1 estimate fare so badly in the risk ratio comparisons? Because of over shrinkage.

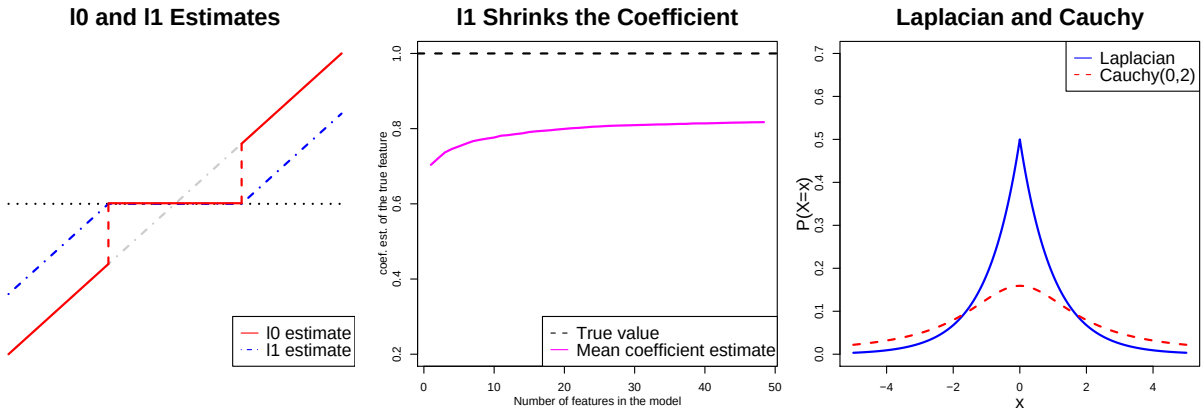

Figure 3: **Left**: The l_0 estimate keeps the least squares value after the cutting point, but the l_1 estimate always shrinks the least squares estimate by a fixed amount. **Middle**: the model we simulate has only one true feature with true $\beta = 1$ and a thousand spurious features. We compute the average Lasso estimate of β for a fixed number of features included in the model (as an index of the l_1 penalty) from several different trials. $\bar{\hat{\beta}}_{l_1}$ is always shrunk by at least 20% in this experiment. **Right**: The Cauchy density has heavier tails than the Laplacian density does. Thus, a Laplacian prior tends to shrink large values of β's.

From a frequentist's point of view, the l_1 estimator (5) shrinks the coefficients and thus is biased (Figure 3). In practice, $\hat{\beta}_{l_1}$ can be substantially shrunk towards zero when the system is sparse, as shown in the middle panel of Figure 3.

From a Bayesian's perspective, the l_1 penalty is equivalent to putting a Laplacian prior on β Tibshirani (1996); Efron et al. (2004), while the l_0 penalty can be approximated by Cauchy priors Johnstone and Silverman (2005); Foster and Stine (2005). The right panel of Figure 3 shows that the Cauchy distribution has a much heavier tail than the Laplacian distribution does. This implies that when the true β is far away from 0, the l_1 penalty will substantially shrink the estimate toward zero.

The bias caused by the shrinkage increases the predictive risk proportionally to the squared amount of the shrinkage. The sparser the problem is, the greater the shrinkage is, thus the larger the risk is.

These results show that in theory the l_0 estimate has a lower risk and provides a more accurate solution. Empirically, stepwise regression performs well in large data sets, where a sparse solution is particularly preferred George and Foster (2000); Foster and Stine (2004); Zhou et al. (2006).

3.3 l_0 controls the False Discovery Rate (FDR) better.

The False Discovery Rate (FDR) is increasingly used to control the fraction of falsely rejected hypotheses (e.g., the number of features added to the model that should not have

been added), especially in fields like biology and genetics, where the interpretation of the data is at least as important as prediction. FDR is defined as $E[V/R|R > 0]P(R > 0)$ Benjamini and Hochberg (1995), where R is the total number of discoveries and V is the number of false discoveries among them. It controls the expected proportion of false positives in multiple testing problems.

The procedure proposed in Abramovich et al. (2006) aims at controlling FDR in models with the assumption of orthogonal predictors. It was shown that this FDR-penalized procedure is adaptively optimal in any l_p ball, $0 \leq p \leq 2$, in the sense of asymptotic minimaxity. The penalty being used is an l_0 type regularization.

To see this difference, we simulated a simple problem for Lasso and stepwise regression to solve. In Table 1, we compare the mean true and false discoveries of coefficients found by forward stepwise regression using RIC and Lasso on synthetic data. The forward stepwise regression does a better job in controlling FDR than Lasso does, when a sparse result is preferred.

	$p = 8$		$p = 1000$		$p = 10000$	
Method	True	False	True	False	True	False
Lasso	4.0	2.37	4.0	26.8	4.0	41.87
Stepwise	4.0	0.61	3.73	0.14	4.0	0.18

Table 1: Mean true and false discoveries of features over 100 tests. In the simulation, the number of effective features is 4 and that of the potential features $p = 8, 1000$ and 10000. The sample size $n = 50$. All the features are independently generated. Lasso tends to have a high FDR when the system is sparse.

4 l_1 optimization using an l_0 criterion

cite Wainright and compare to his results; he notes that under near orthogonal conditions, l_1 gives consistency.

We can make use of the LARS algorithm to generate a set of candidate solutions and then use the l_0 criterion to find the best of the solutions along the regularization path. We evaluated this method as follows. We simulated $\mathbf{y}$ from a thousand features, only 4 of which have nonzero contributions, plus a random noise distributed as $N(0, 1)$. Both the training set and the test set have size $n = 100$. We apply the Lasso algorithm implemented by LARS on this synthetic data set. For each step on the regularization path, this algorithm selects a subset $\mathcal{C} \subset \{1, \ldots, 1000\}$ of features that are included in the model. We then adopt a modified RIC criterion suggested in George and Foster (2000):

$$\|\mathbf{y} - X_C \hat{\beta}_{\mathcal{C}}\|_2^2 + \sum_{q=1}^{|\mathcal{C}|} 2 \log(p/q)\sigma^2 \tag{10}$$

to find an optimal $\mathcal{C}$. The crucial part here is that the coefficient estimate $\hat{\beta}_{\mathcal{C}}$ being used in (10) is the least squares estimate of the true β obtained by fitting $\mathbf{y}$ on $X_{\mathcal{C}} = (\mathbf{x}_j)_{j \in \mathcal{C}}$,

and not the Lasso estimate $\hat{\beta}_{l_1}$ provided by the algorithm. We also use this least squares estimate in out-of-sample calculations.

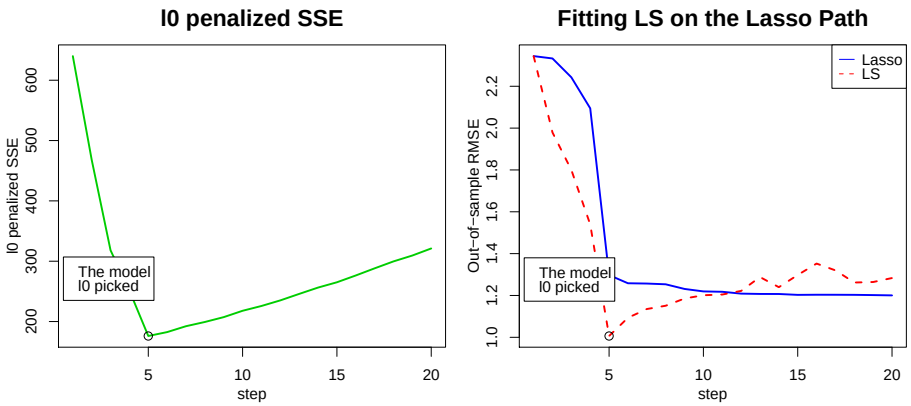

Figure 4: l_0 penalties help finding the best model (independent predictors case). $\mathbf{y}$ is simulated from one thousand features, only four of which have nonzero contributions, plus an $N(0,1)$ error. Both the training set and the test set have sizes $n = 100$. Each step in the LARS algorithm gives a set of features with nonzero coefficient estimates. We compute the least squares (LS) estimates on this subset and the modified RIC criterion (10) on the training set. We also compare the out-of-sample root mean squared errors using the LS estimates and the Lasso estimates on this LARS path. The features are independently generated. The model that minimizes the l_0 penalized error has exactly four variables in it. It also outperforms any of the l_1 models out-of-sample on this data set.

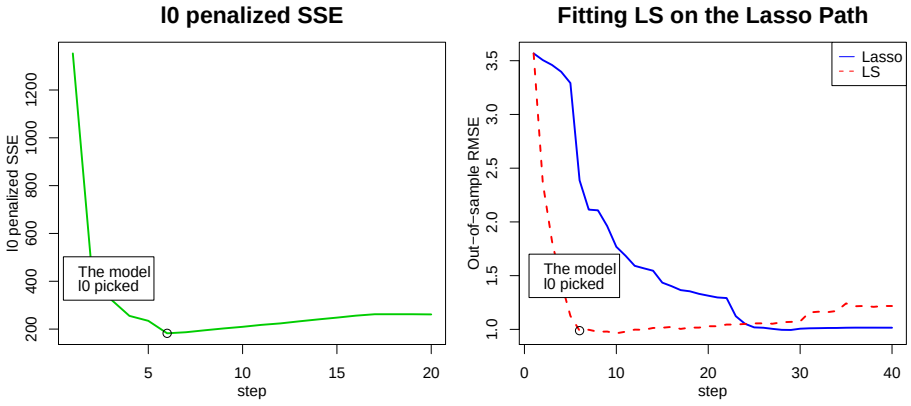

Figure 5: l_0 penalties help finding the best model (correlated predictors case). The setup is exactly the same as in Figure 4 except that each pair of features has a correlation $\rho = 0.64$. In this case, the optimal model under the modified RIC criterion has a slightly better RMSE than the best l_1 model. The Lasso out-of-sample RMSE is typically minimized when the model has included more than 50 features.

We compare two cases: the $\mathbf{x}_j$'s are generated independently of each other, meaning that $X'X$ is diagonal, and the $\mathbf{x}_j$'s are generated with a pairwise correlation $\rho = 0.64$. As shown in Figure 4, in the independent feature case, the model picked by the modified RIC criterion always outperforms any Lasso model on the test set. In the case with correlated predictors (Figure 5), there is little difference between the out-of-sample accuracies of the l_0-picked model and the best Lasso model in this case, but Lasso adds around 50 more spurious variables.

Thus, by combining the computational efficiency of an l_1 algorithm and the sparsity guaranteed by the l_0 penalization, we can easily select an accurate model without cross validation.

Theorem: Under the assumption of incoherence, the risk ratio of l_1 to l_2 when l_1 is followed by ???

Theorem: Under the assumption of incoherence, the l_1 - l_2 risk ratio of obtained when l_1 is used for feature selection, followed by l_0 regression is the same as when l_0 is used.

5 l_0 and NP-hardness

The l_0 problem is NP-hard and hence, at least in theory, intractable. (In practice, of course, people often use approximate solutions to problems that in the worst case can be NP-hard.) One of the attractions of l_1-regularization is that it is convex, hence solvable in polynomial time.

In this section, we compare how the two approaches fare on a known NP-hard regression problem.

We start with a simple constructive proof that the risk ratio for l_1 to l_0 can be arbitrarily bad. Construct data as follows. Pick a large number of independent features z_j. Construct new features $x_1 = z_1 + \epsilon z_2$ and $x_2 = z_1 - \epsilon z_2$ and. Let $y = (z_1 + z_2)/2$ plus noise. Then the correct model is $y = x_2/\epsilon$. Include the rest of the features $z_j, j > 2$ as spurious features. ...

In Natarajan (1995) the known NP hard problem of "the exact cover of 3-sets" was reduced to the best subset selection problem as below: $\mathbf{y} = \mathbf{1}_n$, X is an $n \times p$ binary matrix with each column having three nonzero elements: $\|\mathbf{x}_i\|_0 = 3$, β is a $p \times 1$ vector, $\varepsilon > 0$ and we want to solve

$$\min_{\beta} \|\beta\|_0, \quad \text{s.t.} \quad \|\mathbf{y} - X\beta\|_2 < \varepsilon. \tag{11}$$

Note that if there *is* a solution to this problem, the number of features being chosen should be $n/3$.

We then ask which method comes closer to solving this problem: a greedy approximation to the l_0 problem or an exact solution to the l_1 problem. To this end, we applied Lasso and forward stepwise regression on various n's. For small n's, we took full collections of the three subsets, i.e., p equals n choose 3; for larger n's, we took $p = 10 \cdot n$. Table 2 and 3 list the number of subsets included in the model. Forward stepwise regression always finds fewer subsets, and hence a better solution, than Lasso.

All of our experiments on both synthetic and real data sets show that greedy search algorithms, such as stepwise regression, aimed at minimizing l_0-regularized error provide sparser results. This is because l_0 penalizes the sparsity directly, while l_1 does not. It is

Method	$n = 9$	$n = 12$	$n = 15$	$n = 18$	$n = 21$	$n = 24$	$n = 27$	$n = 30$
Lasso	6	10	11	17	19	21	22	29
Stepwise	3	4	5	6	7	8	9	10

Table 2: The number of subsets chosen by Lasso and by forward stepwise regression with $\varepsilon = 1/4$. All 3-subsets were considered, i.e., $p = \binom{n}{3}$. Forward stepwise regression always has the fewest possible number of subsets, namely, $n/3$.

Method	$n = 99$	$n = 240$	$n = 540$	$n = 990$	$n = 1500$
Lasso	93	219	504	812	1372
	(2×10^{-23})	(9×10^{-23})	(9×10^{-15})	(6×10^{-20})	(2×10^{-20})
Stepwise	40	96	223	364	595
	(1×10^{-28})	(6×10^{-27})	(3×10^{-26})	(6×10^{-25})	(1×10^{-25})

Table 3: The number of subsets chosen by Lasso and by forward stepwise regression with $\varepsilon = 1/4$. $p = 10 \cdot n$ 3-subsets were randomly chosen to be the predictors. Forward stepwise regression always chooses a sparser solution in the sense that it chooses fewer number of subsets. Numbers in parentheses are the sum of squared errors when the algorithms terminated.

easy to construct an example where l_1 will pick a solution with a smaller l_1 norm but with a less sparse solution Candes et al. (2007).

6 Conclusion

REWRITE

Statistically, the l_0 regularization criterion is superior to that of l_1 regularization; l_0 generally provides a more accurate solution and controls the false discovery rate better. l_1 can give arbitrarily worse predictive accuracy than l_0, since l_1 regularization tends to shrink coefficients too much to include many spurious features. Computationally, l_1 appears to be more attractive; convex programming makes the computation feasible and efficient. In practice, however, approximate solutions to the l_0 problem are often better than than exact solutions to the l_1 problem. The best properties of the two methods can be combined. Superior results were obtained by using convex optimization of the l_1 problem to generate a set of candidate models (the regularization path generated by LARS), and then selecting the best model by minimizing the l_0-penalized training error.

7 Appendix: Proofs

We will drop the γ's when the situation is clear, and denote $\hat{\beta}_{l_0}(\gamma_0)$ as $\hat{\beta}_{l_0}$ and $\hat{\beta}_{l_1}(\gamma_1)$ as $\hat{\beta}_{l_1}$ for simplicity.

The l_0 risk can be written as

$$\begin{aligned} R(\beta,\hat{\beta}_{l_0}) &= \mathbb{E}_\beta\|X\beta - X\hat{\beta}\|^2 = \mathbb{E}_\beta\sum_{i=1}^p \|\mathbf{x}_i\|^2(\beta_i-\hat{\beta}_i)^2 \\ &= \mathbb{E}_\beta\sum_{i=1}^p \left((\mathbf{x}_i'\varepsilon/\|\mathbf{x}_i\|)^2 I_{\{|\hat{\beta}_i|>\gamma\}} + (\|\mathbf{x}_i\|\beta_i)^2 I_{\{|\hat{\beta}_i|\le\gamma\}}\right) \qquad (12) \\ &= \sum_{i=1}^p \left\{\sigma^2\mathbb{E}_\beta\left[Z_i^2 I_{\{|\beta_i+\sigma Z_i|>\gamma\}}\right] + (\|\mathbf{x}_i\|\beta_i)^2 P(|\beta_i+\sigma Z_i|\le\gamma)\right\}, \end{aligned}$$

where $Z_i = \mathbf{x}_i'\varepsilon/\sigma\|\mathbf{x}_i\| \sim N(0,1)$, $i = 1,\ldots,p$.

Similarly, the l_1 risk can be written as

$$\begin{aligned} R(\beta,\hat{\beta}_{l_1}) &= \mathbb{E}_\beta\sum_{i=1}^p \Big((\mathbf{x}_i'\varepsilon/\|\mathbf{x}_i\| - \tilde{\gamma})^2 I_{\{\hat{\beta}_i>\tilde{\gamma}\}} + (\mathbf{x}_i'\varepsilon/\|\mathbf{x}_i\| + \tilde{\gamma})^2 I_{\{\hat{\beta}_i<-\tilde{\gamma}\}} \\ &\quad +(\|\mathbf{x}_i\|\beta_i)^2 I_{\{|\hat{\beta}_i|\le\tilde{\gamma}\}}\Big) \qquad (13) \\ &= \sum_{i=1}^p \left\{\mathbb{E}_\beta\left[(\sigma Z_i-\tilde{\gamma})^2 I_{\{\beta_i+\sigma Z_i>\tilde{\gamma}\}} + (\sigma Z_i+\tilde{\gamma})^2 I_{\{\beta_i+\sigma Z_i<-\tilde{\gamma}\}}\right]\right. \\ &\quad \left.+(\|\mathbf{x}_i\|\beta_i)^2 P(|\beta_i+\sigma Z_i|\le\gamma)\right\}, \end{aligned}$$

Without loss of generality, we assume $X'X = I$ and $\sigma = 1$. Specifically, we consider the case when $p = 1$. Let $\Phi(z) = P(Z \le z)$ and $\tilde{\Phi}(z) = P(Z > z)$ be the lower and upper tail probabilities of a standard normal distribution and the two risk functions can be explicitly written as

$$\begin{aligned} R(\beta,\hat{\beta}_{l_0}) &= \int_{\gamma_0-\beta}^{\infty} z^2\phi(z)dz + \int_{-\infty}^{-\gamma_0-\beta} z^2\phi(z)dz + \beta^2\left[\Phi(\gamma_0-\beta) - \tilde{\Phi}(\gamma_0+\beta)\right] \\ &= (\gamma_0-\beta)\phi(\gamma_0-\beta) + (\gamma_0+\beta)\phi(\gamma_0+\beta) \qquad (14) \\ &\quad +\Phi(-\gamma_0+\beta) + \beta^2\Phi(\gamma_0-\beta) + (1-\beta^2)\tilde{\Phi}(\gamma_0+\beta), \\ R(\beta,\hat{\beta}_{l_1}) &= \int_{\gamma_1-\beta}^{\infty}(z-\gamma_1)^2\phi(z)dz + \int_{-\infty}^{-\gamma_1-\beta}(z+\gamma_1)^2\phi(z)dz \\ &\quad +\beta^2\left[\Phi(\gamma_1-\beta) - \tilde{\Phi}(\gamma_1+\beta)\right] \\ &= (-\gamma_1-\beta)\phi(\gamma_1-\beta) + (-\gamma_1+\beta)\phi(\gamma_1+\beta) \qquad (15) \\ &\quad +(\gamma_1^2+1)\Phi(-\gamma_1+\beta) + \beta^2\Phi(\gamma_1-\beta) + (\gamma_1^2+1-\beta^2)\tilde{\Phi}(\gamma_1+\beta). \end{aligned}$$

We list a few Gaussian tail bounds here that we will use in the proofs later. Detailed discussion can be found in related articles Feller (1968); Donoho and Johnstone (1994); Foster and George (1994); Abramovich et al. (2006).

Lemma 3 *For any $z > o$,*

1. $\phi(z)(z^{-1} - z^{-3}) \le \tilde{\Phi}(z) \le \phi(z)z^{-1}$;

2. $\tilde{\Phi}(z) \leq e^{-z^2}/2$.

3. $\phi(z)(x^{-1}-x^{-3}+(1\cdot 3)\cdot x^{-5}-(1\cdot 3\cdot 5)\cdot x^{-7}+\cdots+(-1)^k\cdot(2k-1)!!\cdot x^{-2k-1}$ *overestimates* $\tilde{\Phi}(z)$ *if* k *is even, and underestimates* $\tilde{\Phi}(z)$ *if* k *is odd.*

Lemma 4 *For large enough* $\gamma_0 > 0$,

$$\inf_{\gamma_1}\sup_{\beta}\frac{R(\beta,\hat{\beta}_{l_1})}{R(\beta,\hat{\beta}_{l_0})} > \gamma_0. \tag{16}$$

Proof It suffices to show that for any fixed γ_0 and any γ_1

$$\sup_{\beta}\frac{R(\beta,\hat{\beta}_{l_1})}{R(\beta,\hat{\beta}_{l_0})} > \gamma_0.$$

Suppose $\gamma_1 \geq \gamma_0/\sqrt{2}$, let $\beta_n = (n+1)\gamma_0$, then

$$\|\hat{\beta}_{l_1}-\hat{\beta}_{LS}\|_2^2 > \|\hat{\beta}_{l_1}-\hat{\beta}_{LS}\|_2^2 I_{\{\hat{\beta}_{LS}>\gamma_1\}} = \gamma_1^2 I_{\{\hat{\beta}_{LS}>\gamma_1\}} \geq \frac{\gamma_0^2}{2} I_{\{\hat{\beta}_{LS}>\gamma_1\}}.$$

Hence,

$$\mathbb{E}\|\hat{\beta}_{l_1}-\hat{\beta}_{LS}\|_2^2 > \frac{\gamma_0^2}{2}P(\hat{\beta}_{LS}>\gamma_1),$$

where $Z \sim N(0,1)$. Thus,

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}_{l_1}-\beta_n\|_2^2 &\geq E\|\hat{\beta}_{l_1}-\hat{\beta}_{LS}\|_2^2 - E\|\hat{\beta}_{LS}-\beta_n\|_2^2 > \frac{\gamma_0^2}{2}P(\hat{\beta}_{LS}>\gamma_1) - 1 \\
&= \left(\frac{\gamma_0^2}{2}-1\right)P(\hat{\beta}_{LS}>\gamma_1) - P(\hat{\beta}_{LS}\leq\gamma_1) \\
&> \gamma_0\Phi((n+1)\gamma_0-\gamma_1) - \Phi(\gamma_1-(n+1)\gamma_0),
\end{aligned}$$

for large enough γ_0.

On the other hand,

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}_{l_0}-\beta_n\|_2^2 &= -n\gamma_0\phi(n\gamma_0) + (n+2)\gamma_0\phi\left((n+2)\gamma_0\right) + \Phi(n\gamma_0) \\
&\quad +(n+1)^2\gamma_0^2\tilde{\Phi}(n\gamma_0) + (1-(n+1)^2\gamma_0^2)\tilde{\Phi}((n+2)\gamma_0) \\
&\leq 1 + \left(-n\gamma_0 - \frac{1}{2n\gamma_0} + \frac{(n+1)^2\gamma_0}{n}\right)\phi(n\gamma_0) \\
&\quad + \left((n+2)\gamma_0 + \frac{1-(n+1)^2\gamma_0^2}{(n+2)\gamma_0}\right)\phi((n+2)\gamma_0) \\
&\leq 1 + \left(2 + \frac{1}{n} + 2e^{-2(n+1)\gamma_0^2}\right)\gamma_0\phi(n\gamma_0).
\end{aligned}$$

Hence,

$$\frac{R(\beta_n, \hat{\beta}_{l_1})}{R(\beta_n, \hat{\beta}_{l_0})} \geq \frac{\gamma_0 \Phi((n+1)\gamma_0 - \gamma_1) - \Phi(\gamma_1 - (n+1)\gamma_0)}{1 + \left(2 + n^{-1} + 2e^{-2(n+1)\gamma_0^2}\right)\gamma_0 \phi(n\gamma_0)}$$

Let $n \to \infty$, then

$$\sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_1})}{R(\beta, \hat{\beta}_{l_0})} \geq \lim_{n\to\infty} \frac{R(\beta_n, \hat{\beta}_{l_1})}{R(\beta_n, \hat{\beta}_{l_0})} \geq \gamma_0. \tag{17}$$

For those $0 \leq \gamma_1 < \gamma_0/\sqrt{2}$, we consider $\beta = 0$ and denote

$$\begin{aligned} R_0(\gamma_0) &= R(0, \hat{\beta}_{l_1}(\gamma_1)) = 2\gamma_0\phi(\gamma_0) + 2\tilde{\Phi}(\gamma_0) & (18)\\ R_1(\gamma_1) &= R(0, \hat{\beta}_{l_0}(\gamma_0)) = -2\gamma_1\phi(\gamma_1) + 2(\gamma_1^2+1)\tilde{\Phi}(\gamma_1). & (19) \end{aligned}$$

We first show that for $c \leq \gamma_1 < \gamma_0/\sqrt{2}$, $R_1(\gamma_1)/R_0(\gamma_0) > \gamma_0$, where c is a constant such that $\tilde{\Phi}(z) - \phi(z)(1/z - 1/z^3 + 1/z^5) \geq 0$, for any $z \geq c$, then since

$$\frac{d}{d\gamma_1} R_1(\gamma_1) = -4\phi(\gamma_1) + 4\gamma_1\tilde{\Phi}(\gamma_1) < 0,$$

for any $0 \leq \gamma_1 \leq c$

$$\frac{R_1(\gamma_1)}{R_0(\gamma_0)} \geq \frac{R_1(c)}{R_0(\gamma_0)} > \gamma_0.$$

For any $c \leq \gamma_1 < \gamma_0/\sqrt{2}$, we have $\phi(\gamma_1) \geq \phi(\gamma_0)e^{-\gamma_0^2/4}$, and

$$\begin{aligned} R_1(\gamma_1) &\geq -2\gamma_1\phi(\gamma_1) + 2(\gamma_1^2+1)\left(\gamma_1^{-1} - \gamma_1^{-3} + \gamma_1^{-5}\right)\phi(\gamma_1)\\ &\geq 2\gamma_1^{-5}\phi(\gamma_1) \geq 2^{7/2}\gamma_0^{-5}\phi(\gamma_0)e^{\gamma_0^2/4}\\ R_0(\gamma_0) &\leq 2\left(\gamma_0 + \gamma_0^{-1}\right)\phi(\gamma_0). \end{aligned}$$

Thus for large enough γ_0

$$\frac{R_1(\gamma_1)}{R_0(\gamma_0)} \geq \frac{2^{5/2}e^{\gamma_0^2/4}}{\gamma_0^6 + \gamma_0^4} > \gamma_0.$$

■

Proof of Theorem 1: Let $C_1 = -\arg\min_z\{z > 2 : 2^{5/2}e^{z^2/4} - z^7 - z^5 > 0\}$. □

Lemma 5 *There exists an $M > 0$ and a constant $C > 0$, such that for all $\gamma_1 > M$, $\gamma_0 \equiv \gamma_1 + 4\log\gamma_1/\gamma_1$,*

$$\sup_{\beta} \frac{R(\beta, \hat{\beta}_{l_0})}{R(\beta, \hat{\beta}_{l_1})} \leq 1 + C \cdot \gamma_1^{-1}. \tag{20}$$

Proof

It suffices to show that for all $\beta \geq 0$, we have

$$\frac{R(\beta, \hat{\beta}_{l_0})}{R(\beta, \hat{\beta}_{l_1})} \leq 1 + C \cdot \gamma_1^{-1}. \tag{21}$$

The proof is done by generating bounds for the risks at various β's.

We first show that this is true when $0 \leq \beta < \log \gamma_1/\gamma_1$. When $\beta \approx 0$, the left hand side of (21) is dominated by rejecting parts. We thus have

$$R(\beta, \hat{\beta}_1) \geq \frac{2\sqrt{\log \gamma_1}}{\gamma_1} P\left(Z \geq \gamma_1 + \frac{2\sqrt{\log \gamma_1}}{\gamma_1}\right) \tag{22}$$

and

$$R(\beta, \hat{\beta}_0) \leq 2\gamma_1^2 P\left(Z \geq \gamma_1 + \frac{6\sqrt{\log \gamma_1}}{\gamma_1}\right) \tag{23}$$

The $P()$ term in (23) goes to zero much faster than the $P()$ term in (22). Hence the ratio converges to zero for large γ_1.

For small β, namely $|\beta| \leq \gamma_1 - 2\sqrt{\log \gamma_1}$ we have

$$R(\beta, \hat{\beta}_1) \geq \beta^2 P(|Z| \leq \gamma_1 - |\beta|), \tag{24}$$

and

$$R(\beta, \hat{\beta}_0) \leq \beta^2 P(|Z| \leq \gamma_1 - |\beta|) + 2\gamma_1 P(|Z| \geq \gamma_0 - \beta). \tag{25}$$

The two risks are close to each other relative to the size of the l_1 risk.

For $|\beta| \geq \gamma_1 + 1$ we have

$$R(\beta, \hat{\beta}_1) \geq R(\beta, \hat{\beta}_0). \tag{26}$$

This is because

$$\begin{aligned}
R(\beta, \hat{\beta}_{l_0}) &= (\gamma_0 - \beta)\phi(\gamma_0 - \beta) + (\gamma_0 + \beta)\phi(\gamma_0 + \beta) \\
&\quad + \Phi(-\gamma_0 + \beta) + \beta^2\Phi(\gamma_0 - \beta) + (1 - \beta^2)\tilde{\Phi}(\gamma_0 + \beta) \\
&= (\gamma_1 + \Delta\gamma - \beta)\phi(\gamma_1 - \beta) + (\gamma_1 + \Delta\gamma - \beta)\left.\frac{\partial}{\partial\gamma}\phi(\gamma - \beta)\right|_{\gamma_1}\Delta\gamma \\
&\quad + (\gamma_1 + \Delta\gamma + \beta)\phi(\gamma_1 + \beta) + (\gamma_1 + \Delta\gamma + \beta)\left.\frac{\partial}{\partial\gamma}\phi(\gamma + \beta)\right|_{\gamma_1}\Delta\gamma \\
&\quad + \Phi(-\gamma_1 + \beta) + \left.\frac{\partial}{\partial\gamma}\Phi(-\gamma + \beta)\right|_{\gamma_1}\Delta\gamma + \beta^2\Phi(\gamma_1 - \beta) + \beta^2\left.\frac{\partial}{\partial\gamma}\Phi(\gamma - \beta)\right|_{\gamma_1}\Delta\gamma \\
&\quad + (1 - \beta^2)\tilde{\Phi}(\gamma_1 + \beta) + (1 - \beta^2)\left.\frac{\partial}{\partial\gamma}\tilde{\Phi}(\gamma + \beta)\right|_{\gamma_1}\Delta\gamma + \gamma_1 e^{-\gamma_1^2/2}o(\Delta\gamma) \\
&= (\gamma_1 - \beta)\phi(\gamma_1 - \beta) + (\gamma_1 + \beta)\phi(\gamma_1 + \beta) + \Phi(-\gamma_1 + \beta) + \beta^2\Phi(\gamma_1 - \beta) \\
&\quad + (1 - \beta^2)\tilde{\Phi}(\gamma_1 + \beta) - (\gamma_1^2 - 2\beta\gamma_1)\phi(\gamma_1 - \beta)\Delta\gamma \\
&\quad - (\gamma_1^2 + 2\beta\gamma_1)\phi(\gamma_1 + \beta)\Delta\gamma + \gamma_1 e^{-\gamma_1^2}o(\Delta\gamma) \\
&= R(\beta, \hat{\beta}_{l_1}) + 2\gamma_1\phi(\gamma_1 - \beta) + 2\gamma_1\phi(\gamma_1 + \beta) - \gamma_1^2\Phi(-\gamma_1 + \beta) - \gamma_1^2\tilde{\Phi}(\gamma_1 + \beta) \\
&\quad - (\gamma_1^2 - 2\beta\gamma_1)\phi(\gamma_1 - \beta)\Delta\gamma - (\gamma_1^2 + 2\beta\gamma_1)\phi(\gamma_1 + \beta)\Delta\gamma + \gamma_1 e^{-\gamma_1^2/2}o(\Delta\gamma)
\end{aligned}$$

We consider $\Delta\gamma = 4\log(\gamma_1)/\gamma_1$. When $\beta \geq \gamma_1 + 1$, and γ_1 is large, we have

$$\begin{aligned} & R(\beta, \hat{\beta}_{l_0}) - R(\beta, \hat{\beta}_{l_1}) \\ \leq & -\gamma_1^2 + (2\gamma_1 - (\beta - \gamma_1)^{-1}/2 - 4\gamma_1 \log(\gamma_1) + 8\beta \log(\gamma_1))\phi(\gamma_1 - \beta) \\ & +(2\gamma_1 - (\beta + \gamma_1)^{-1}/2 - 4\gamma_1 \log(\gamma_1) - 8\beta \log(\gamma_1))\phi(\gamma_1 + \beta) + \gamma_1 e^{-\gamma_1^2/2} o(\Delta\gamma) \\ < & 0. \end{aligned}$$

For β close to γ_1, namely $\gamma_1 - 2\sqrt{\log \gamma_1} \leq |\beta| \leq \gamma_1 + 1$ we have

$$R(\beta, \hat{\beta}_1) \geq \gamma_1^2/2 \tag{27}$$

and

$$\begin{aligned} R(\beta, \hat{\beta}_1) - R(\beta, \hat{\beta}_0) \leq & \gamma_1 P\left(\gamma_1 - \beta \leq Z \leq \gamma_1 - \beta + \frac{8\sqrt{\log \gamma_1}}{\gamma_1}\right) \\ & -\frac{\gamma_1^2}{4} P\left(\gamma_1 - \beta + \frac{8\sqrt{\log \gamma_1}}{\gamma_1} \leq Z \leq \frac{9}{8}\gamma_1 - \beta\right) + \gamma_1^2 e^{-\gamma_1/8} \end{aligned}$$

This case is where the C in (20) gets its actual value. We have made this gap more explicit. Our result follows from computing with the actual tail bounds for the normal distribution. The proof of this penultimate case requires using Lemma 6.

■

Lemma 6 *For $X = \beta + Z$ where $Z \sim N(0,1)$, and $\hat{\beta} = f(X)$ with $|\hat{\beta} - X| < \gamma$,*

$$\mathbb{E}\left[\left(\hat{\beta} - \beta\right)^2 I_A\right] \leq (\gamma + 2)^2 \sqrt{P(A)}$$

Proof

$$\begin{aligned} \mathbb{E}\left[\left(\hat{\beta} - \beta\right)^2 I_A\right] = & \mathbb{E}\left[\left(\hat{\beta} - X + X - \beta\right)^2 I_A\right] \leq \mathbb{E}\left[\left(|\hat{\beta} - X| + |Z|\right)^2 I_A\right] \\ \leq & \mathbb{E}\left[(\gamma + |Z|)^2 I_A\right] \leq \sqrt{\mathbb{E}\left[(\gamma + |Z|)^4\right] \mathbb{E}[I_A^2]} \end{aligned}$$

Where

$$\begin{aligned} \mathbb{E}\left[(\gamma + |Z|)^4\right] = & \gamma^4 + 4\gamma^3 \mathbb{E}|Z| + 6\gamma^2 \mathbb{E}Z^2 + 4\gamma \mathbb{E}|Z|^3 + \mathbb{E}Z^4 \\ \leq & \gamma^4 + 4\gamma^3 2 + 6\gamma^2 4 + 4\gamma 8 + 16 = (\gamma + 2)^4 \end{aligned}$$

■

Proof of Theorem 2: For $\gamma_1 < M$ we know that there exists some $\epsilon > 0$ such that $R(\beta, \hat{\beta}_{l_1}(\gamma_1)) \geq \epsilon$ for all β. If we use the trivial estimator $\gamma_0 = 0$, we know it has risk 1. Hence, we can pick $C_2 = \max(1/\epsilon, C)$ where C is from our lemma, then Theorem 2 follows. □

References

F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281, 1973.

M. M. Barbieri and J. O. Berger. Optimal Predictive Model Selection. *The Annals of Statistics*, 32(3):870–897, 2004.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Ser. B*, 57(1):289–300, 1995.

E. J. Candes and T. Tao. The Dantzig Selector: Statistical Estimation When p is much larger than n. *The Annals of Statistics*, 35(6), 2007.

E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing Sparsity by Reweighted l_1 Minimization, 2007.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression (with discussion). *The Annals of Statistics*, 32(2), 2004.

W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley, January 1968. ISBN 0471257087.

D. P. Foster and E. I. George. The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

D. P. Foster and R. A. Stine. Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99(466):303–313, 2004.

D. P. Foster and R. A. Stine. Polyshrink: An adaptive variable selection procedure that is competitive with Bayes experts. Technical report, Univ. of Penn., 2005.

E. I. George and D. P. Foster. Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87(4):731–747, 2000.

I. M. Johnstone and B. W. Silverman. Empirical Bayes Selection of Wavelet Thresholds. *The Annals of Statistics*, 33(4):1700–1752, 2005.

D. Lin, D. P. Foster, E. Pitler, and L. H. Ungar. In Defense of l_0 Regularization, 2008.

N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Ser. B*, 67(1):91–108, 2005.

M. Yuan and Y. Lin. Model Selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Ser. B*, 68:49–67, 2006.

J. Zhou, D. P. Foster, R. A. Stine, and L. H. Ungar. Streamwise Feature Selection. *Journal of Machine Learning Research*, 7:1861–1885, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B*, 67:301–320, 2005.