On the Sensitivity of the Lasso to the Number of Predictor Variables

Cheryl J. Flynn Clifford M. Hurvich Jeffrey S. Simonoff *

Abstract
The Lasso is a computationally efficient procedure that can produce sparse estimators when the number of predictors (p) is large. Oracle inequalities provide probability loss bounds for the Lasso estimator at a deterministic choice of the regularization parameter. These bounds tend to zero if p is appropriately controlled, and are thus commonly cited as theoretical justification for the Lasso and its ability to handle high-dimensional settings. Unfortunately, in practice the regularization parameter is not selected to be a deterministic quantity, but is instead chosen using a random, data-dependent procedure. To address this shortcoming of previous theoretical work, we study the loss of the Lasso estimator when tuned optimally for prediction. Assuming orthonormal predictors and a sparse true model, we prove that the best possible predictive performance of the Lasso deteriorates as p increases with positive probability. We further demonstrate empirically that the deterioration in performance can be far worse than suggested by the commonly held views in the literature and that this deterioration persists as the sample size increases.

Key Words: Least Absolute Shrinkage and Selection Operator (Lasso), Oracle Inequalities, High-Dimensional Data

1. Introduction

Regularization methods perform model selection subject to the choice of a regularization parameter, and are commonly used when the number of predictor variables is too large to consider all subsets. In regularized regression, these methods operate by minimizing the penalized least squares function

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda \sum_{j=1}^{p} p(\beta_j)$$

where $\boldsymbol{y}$ is a $n \times 1$ response vector, $\boldsymbol{X}$ is a $n \times p$ deterministic matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients, and $p(\cdot)$ is a penalty function. A common choice for the penalty function is the L_1 norm of the coefficients. This penalty function was first proposed by Tibshirani (1996) and termed the Lasso (Least absolute shrinkage and selection operator). The solution to the Lasso is sparse in that it automatically sets some coefficients equal to zero, and the entire regularization path can be found using the computationally efficient Lars algorithm (Efron et al., 2004). Given its computational advantages, understanding the theoretical properties of the Lasso is an important research area.

This paper focuses on the performance of the Lasso for predictive purposes. To that end, we evaluate the Lasso estimated models using the L_2 loss function. Assume that the true data generating process is

$$\boldsymbol{y} = \boldsymbol{\mu} + \varepsilon \tag{1.1}$$

*Leonard N. Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012

where $\boldsymbol{\mu}$ is a $n \times 1$ unknown mean vector and ε is a $n \times 1$ random noise vector. The L_2 loss is defined as

$$L(\lambda) = \frac{||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}||^2}{n} = \frac{||\boldsymbol{\mu} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda||^2}{n} \tag{1.2}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the Lasso estimated vector of coefficients for a specific choice of the regularization parameter λ and $|| \cdot ||^2$ is the squared Euclidean norm. If the true model is included amongst the candidate models, then $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}_0$ for some unknown true coefficient vector $\boldsymbol{\beta}_0$ and the L_2 loss function takes the form

$$L(\lambda) = \frac{||\boldsymbol{X}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_\lambda)||^2}{n}.$$

In most modern applications, it is assumed that $\boldsymbol{\beta}_0$ is sparse and only has $p_0 < p$ non-zero entries.

A large portion of the regularization literature has focused on establishing probability loss bounds for the Lasso and its variants (see e.g., Bunea et al., 2007, Negahban et al., 2009, Bickel et al., 2010, and Buhlmann and van de Geer, 2011). Roughly, for a deterministic choice of λ, these probability bounds are of the form

$$L(\lambda) \leq k\sigma^2 \frac{\log(p)p_0}{n} \tag{1.3}$$

(Buhlmann and van de Geer, 2011). Here σ^2 is the true error variance, and k is a constant that does not depend on p or n.

These bounds are commonly termed "oracle inequalities" since, apart from the $\log(p)$ term and the constant, they closely resemble the loss expected if an oracle told us the true set of predictors and we fit least squares. In light of this connection, it is commonly noted in the literature that the "l_1-estimator achieves the ideal risk ... up to a logarithmic $\log(p)$" (Fan and Lv, 2008) and that the "[$\log(p)$ factor] can be seen as the price to pay for not knowing the active set"(p. 102, Buhlmann and van de Geer, 2011). Furthermore, since the bound depends on n, one can conclude that "the ambient dimension and structural parameters can grow as some function of the sample size n, while still having the statistical error decrease to zero" (Negahban et al., 2009). Similar asymptotic conclusions exist in the work by Greenshtein and Ritov (2004) on the "persistence" of the Lasso estimators. In this context, the authors showed that the difference in the expected prediction error of the Lasso estimator and the optimal estimator converges to zero in probability. From this result, the authors concluded that there "there is 'asymptotically no harm' in introducing many more explanatory variables than observations." The extended work by Greenshtein (2006) similarly concludes that "in some 'asymptotic sense', when assuming a sparsity condition, there is no loss in letting [p] be much larger than n."

Unfortunately, there is a disconnect between these oracle inequalities and the way that the Lasso is implemented in practice. In practice λ is not taken to be a deterministic value, but rather it is selected using an information criterion such as Akaike's information criterion (AIC; Akaike, 1973), the corrected AIC (AIC_c; Hurvich and Tsai, 1989), the Bayesian information criterion (BIC; Schwarz, 1978), or Generalized cross-validation (GCV; Craven and Wahba, 1978) or using the data-dependent procedure 10-fold cross-validation (CV) (see, e.g., Fan and Li, 2001, Leng et al., 2006, Zou et al., 2007, and Flynn et al., 2013). This motivates us to study the behavior of the loss based on a data-dependent choice of the tuning parameter.

Define the random variable $\lambda^* = \text{argmin}_\lambda L(\hat{\boldsymbol{\beta}}_\lambda)$ to be the optimal (infeasible) choice of λ that minimizes the loss function. In what follows, we will focus on the loss of the Lasso evaluated at λ^*. By the definition of λ^*, the loss bound (1.3) still applies, making it possible to compare the observed performance against the commonly held views in the literature. Furthermore, although this choice of the regularization parameter is infeasible, it is the ultimate goal for any model selection procedure and Flynn et al. (2013) showed that the loss at λ selected by 10-fold CV or AIC_c is close to optimal.

The remainder of this paper is organized as follows. Section 2 presents some theoretical results on the behavior of the Lasso and proves that the best case predictive performance can deteriorate as the number of predictor variables is increased. Section 3 investigates the rate of deterioration empirically and shows that it can be much worse than one would expect based on the established loss bounds. Finally, Section 4 presents some final remarks and areas for future research. The appendix includes some additional technical results.

2. Theoretical Results

Here we consider a simple framework for which there exists an exact solution for the Lasso estimator. We assume that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \varepsilon$$

where $\boldsymbol{y}$ is the $n \times 1$ response vector, $\boldsymbol{X}$ is a $n \times p$ matrix of determinisitic predictors such that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ (the $p \times p$ identity matrix), $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_p)^T$ is the $p \times 1$ vector of true unknown coefficients, and $\boldsymbol{\varepsilon}$ is a $n \times 1$ noise vector where $\varepsilon_i \sim_{iid} N(0, \sigma^2)$. For simplicity, we will assume that $\beta_1 \neq 0$ but all of the other true coefficients are equal to zero.

By construction $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{y}$ is the vector of the least squares estimated coefficients based on the full model. It follows that

$$z_1 \sim N(\beta_1, \sigma^2)$$

and

$$z_j \sim_{iid} N(0, \sigma^2)$$

for $2 \leq j \leq p$, and that z_1 is independent of z_j for $2 \leq j \leq p$. Furthermore, for a given λ, the Lasso estimated coefficients are

$$\hat{\beta}_{\lambda j} = \text{sgn}(z_j)(|z_j| - \lambda)_+$$

for $j = 1, \ldots, p$ (Fan and Li, 2001). To measure the performance of this estimator, define

$$L_p(\lambda) = \frac{||\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda||^2}{n}.$$

Here we subscript the loss by p in order to emphasize that the loss at a particular value of λ depends on the number of predictor variables. In particular, for this example,

$$L_p(\lambda) = \frac{1}{n}(\beta_1 - \hat{\beta}_{\lambda 1})^2 + \frac{1}{n}\sum_{j=2}^{p} \hat{\beta}_{\lambda j}^2. \tag{2.1}$$

We wish to study the sensitivity of the Lasso to the number of predictor variables. To do this, we'll vary the number of predictor variables in a nested fashion, so that if $p_1 < p_2$, then the p_1 predictors are a subset of the p_2 predictors. Under this set-up, if $p_1 < p_2$, then $L_{p_1}(\lambda) \leq L_{p_2}(\lambda)$ for any given λ. In what follows, we establish the stronger result that $\min_\lambda L_{p_1}(\lambda) < \min_\lambda L_{p_2}(\lambda)$ with non-zero probability.

Define

$$\lambda^*_{p_1} = \underset{\lambda}{\operatorname{argmin}}\, L_{p_1}(\lambda) \qquad \text{and} \qquad \lambda^*_{p_2} = \underset{\lambda}{\operatorname{argmin}}\, L_{p_2}(\lambda).$$

Under the orthonormality assumption, we require $p_1, p_2 \leq n$.

Theorem 2.1. *If* $1 \leq p_1 < p_2$*, then*

$$\Pr\left(L_{p_2}(\lambda^*_{p_2}) > L_{p_1}(\lambda^*_{p_1})\right) > 0.$$

The proof of Theorem 2.1 makes use of the following two lemmas.

Lemma 2.2. *Let* $1 < p_1$*. For any* $0 < \varepsilon < |\beta_1|$*, if* $z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon]$ *and*

$$\max_{2 \leq j \leq p_1} |z_j| < \left(\frac{\beta_1^2 - \varepsilon^2}{p_1}\right)^{1/2}, \tag{2.2}$$

then $L_{p_1}(0) < \frac{1}{n}\beta_1^2$*.*

Lemma 2.3. *Let* $1 \leq p_1 < p_2$*. For any* $0 < \varepsilon < |\beta_1|$*, if* $z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon]$ *and*

$$|\beta_1| + \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|) < \max_{p_1+1 \leq j \leq p_2} |z_j|, \tag{2.3}$$

then $L_{p_2}(\lambda^*_{p_2}) = \frac{1}{n}\beta_1^2$*.*

The proofs of Lemmas 2.2 and 2.3 are presented in Appendix A. Next, we prove our main result.

Proof of Theorem 2.1. Consider

$$\Pr(L_{p_2}(\lambda^*_{p_2}) > L_{p_1}(\lambda^*_{p_1})) \geq \Pr\left(L_{p_1}(0) < \frac{1}{n}\beta_1^2, L_{p_2}(\lambda^*_{p_2}) = \frac{1}{n}\beta_1^2\right).$$

If $p_1 > 1$, then by Lemmas 2.2 and 2.3, for any $0 < \varepsilon < |\beta_1|$,

$$\begin{aligned}\Pr\left(L_{p_2}(\lambda^*_{p_2}) = \frac{1}{n}\beta_1^2, L_{p_1}(0) < \frac{1}{n}\beta_1^2\right) \geq \Pr\Big(&z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon],\\ &\max_{2 \leq j \leq p_1} |z_j| < \left(\frac{\beta_1^2 - \varepsilon^2}{p_1}\right)^{1/2},\\ &\max_{p_1+1 \leq j \leq p_2} |z_j| > |\beta_1| + \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|)\Big).\end{aligned}$$

Since the $z_j's$, $1 \leq j \leq p_2$, are independent normals, β_1 and ε are fixed constants, and $\left(\frac{\beta_1^2-\varepsilon^2}{p_1}\right)^{1/2}$ and $|\beta_1| + \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|)$ are strictly positive, this probability is strictly positive.

Similarly, if $p_1 = 1$, then by Lemma 2.3, for any $0 < \varepsilon < |\beta_1|$,

$$\begin{aligned}\Pr\left(L_{p_2}(\lambda^*_{p_2}) = \frac{1}{n}\beta_1^2, L_{p_1}(0) < \frac{1}{n}\beta_1^2\right) \geq \Pr\Big(&z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon],\\ &\max_{p_1+1 \leq j \leq p_2} |z_j| > |\beta_1| + \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|)\Big).\end{aligned}$$

This probability is also strictly positive.

It follows then that

$$\Pr(L_{p_2}(\lambda^*_{p_2}) > L_{p_1}(\lambda^*_{p_1})) > 0.$$

□

As a final remark, note that as p_2 gets farther away from p_1, the probability that $\max_{p_1+1 \leq j \leq p_2} |z_j| > |\beta_1| + \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|)$ increases, so the lower bound on this probability increases.

Theorem 2.1 establishes that, with non-zero probability, the best case predictive performance for the Lasso deteriorates as more predictors with no predictive power are added to a given set of predictors. Intuitively, this conclusion makes sense. Increasing the number of noisy predictor variables will make more shrinkage optimal, and more shrinkage results in more bias. However, it is important to note that this cannot happen if one uses least squares with all subsets. In that case the optimal loss would never increase as the number of predictor variables increases. Thus, we are paying a price for not knowing the true set of predictor variables. The next section investigates this price.

3. Empirical Study

This section investigates the cost of not knowing the true set of predictors when working with high dimensional data. We assume that $\boldsymbol{y}$ is generated according to the generating model in (1.1). We consider two simulation set-ups. The first is in line with our theoretical work and studies the performance of the Lasso when the columns of $\boldsymbol{X}$ are trigonometric predictors. Since these predictors are orthogonal, this setting requires $p < n$. To allow for situations with $p > n$, we also study the case where the columns of $\boldsymbol{X}$ are independent standard normals.

The main goal of our simulations is to understand the behavior of the infeasible optimal loss for the Lasso as p and n vary. We focus on cases where p is large or is getting large in order to be consistent with high-dimensional frameworks.

To measure the deterioration in optimal loss we consider the optimal loss ratio

$$\frac{\min_\lambda L_{p_2}(\lambda)}{\min_\lambda L_{p_1}(\lambda)},$$

which compares the minimum loss based on p_2 predictors to the minimum loss based on p_1 predictors. Here $p_1 < p_2$ and the p_1 predictors are a subset of the p_2 predictors.

3.1 Orthogonal Predictors

Define the true model to be

$$y_i = 6x_{i,1} + 5x_{i,2} + 4x_{i,3} + 3x_{i,4} + 2x_{i,5} + x_{i,6} + \varepsilon_i \tag{3.1}$$

for $i = 1, \ldots, (n-1)$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We compare $\sigma^2 = 4$ and $\sigma^2 = 400$ in order to study the impact of varying the signal-to-noise ratio (SNR). We refer to these cases as “High SNR” and “Low SNR”, respectively.

The columns of $\boldsymbol{X}$ are trigonometric predictors defined by

$$x_{i,2j-1} = \sin\left(\frac{2\pi j}{n} i\right)$$

and,

$$x_{i,2j} = \cos\left(\frac{2\pi j}{n} i\right)$$

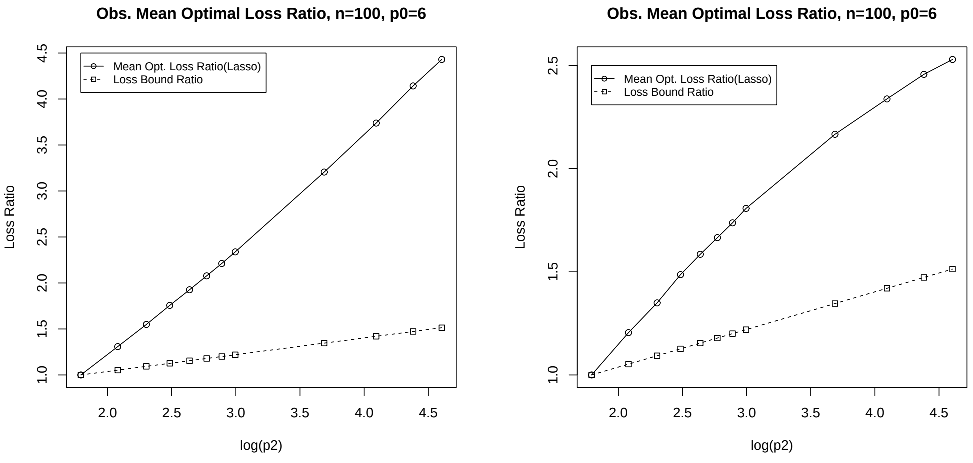
for $j = 1, \ldots, p/2$ and $i = 0, \ldots, n-1$. The columns of $\boldsymbol{X}$ are orthogonal under this design and the true model is always included amongst the candidate models.

By the definition of the optimal loss, the oracle inequalities in the literature also apply to $\min_\lambda L_p(\lambda)$. In particular, applying Corollary 6.2 in Buhlmann and van de Geer (2011), it follows that

$$\min_\lambda L_p(\lambda) \leq 64\sigma^2 p_0 \frac{t^2 + 2\log(p)}{n\psi_0^2} \tag{3.2}$$

with probability greater than $1 - 2e^{-t^2/2}$, where ψ_0 is a constant that satisfies a compatibility condition. This condition places a restriction on the minimum eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}/n$ for a restricted set of coefficients and it's sufficient to take $\psi_0 = 1$ for an orthogonal design matrix. Unless noted otherwise, t is set so that the bound holds with 95 percent probability. Since these bounds also depend on p, we study if the deterioration in optimal loss is adequately predicted by these bounds. In other words, is the price that we pay equal to $\log(p)$?

Figure 1: Mean optimal loss ratio over 1000 realizations as a function of $\log(p_2)$ for $n = 100$ and $p_1 = 6$. The number of predictor variables p_2 is varied from 6 to 100. The "High SNR" and "Low SNR" settings correspond to $\sigma^2 = 4$ and $\sigma^2 = 400$, respectively.

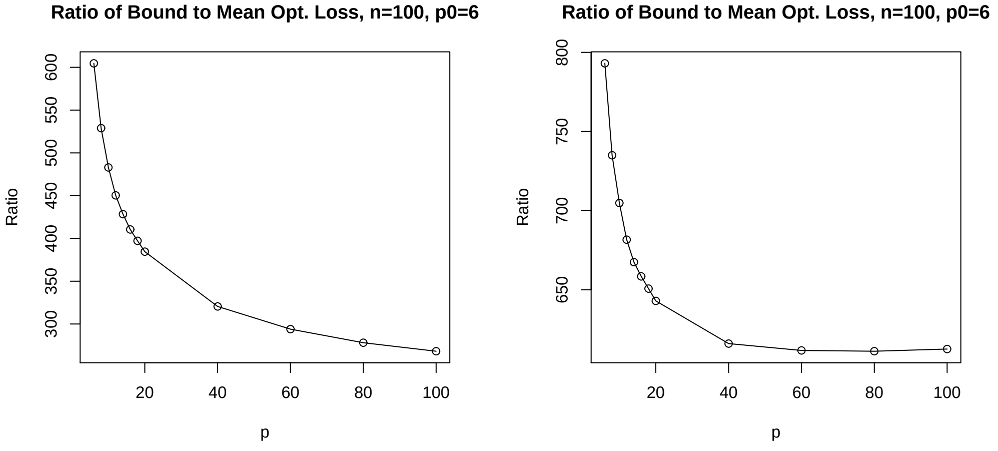


(a) High SNR

(b) Low SNR

Figure 1 compares the means of optimal loss ratios over 1000 realizations to the ratio predicted by the loss bound. We set p_1 equal to the true set of six predictors and vary p_2. The bottom line is the optimal loss ratio predicted by the bound, whereas the top line is the observed mean optimal loss ratios. Clearly the deterioration is far worse than predicted by the bound. For example, if we include n predictors, then the loss bound suggests we should be about 50% worse off than if we knew the true set of predictors, but in actuality we are about 300% worse off on average. This discrepancy is a consequence of the fact that the bounds are inequalities not equalities.

Figure 2: Ratio of the loss bound to the observed mean optimal loss over 1000 realizations as a function of p for $n = 100$. The "High SNR" and "Low SNR" settings correspond to $\sigma^2 = 4$ and $\sigma^2 = 400$, respectively.

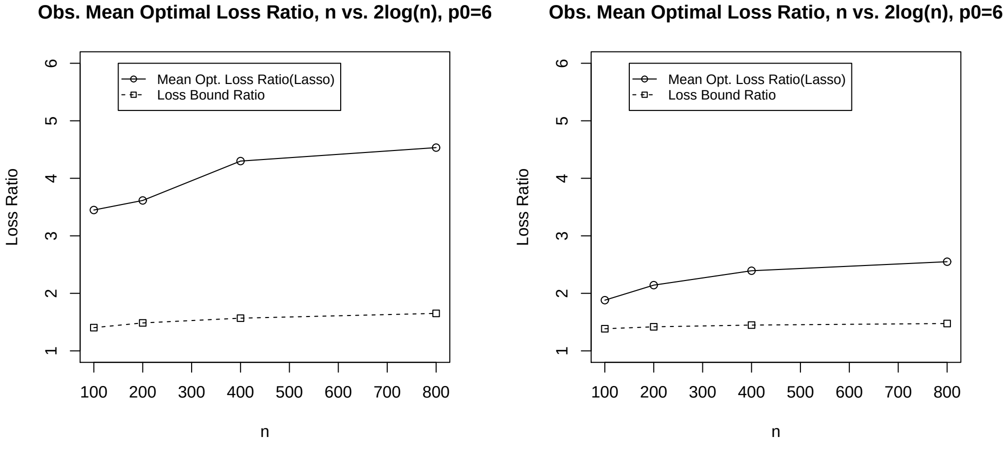


(a) High SNR

(b) Low SNR

To emphasize the danger of relying only on bounds, Figure 2 plots the ratio of the bound to the mean optimal loss for varying values of p. This plot suggests that the bounds are overly conservative when compared to the optimal loss and the degree of conservatism depends on both p and the signal-to-noise ratio. As a result of this behavior, the deterioration in optimal loss can be much worse than $\log(p)$.

Figure 3: Mean optimal loss ratio for $p_2 = n$ predictors compared to $p_1 = 2\log(n)$ predictors over 1000 realizations as a function of n. The "High SNR" and "Low SNR" settings correspond to $\sigma^2 = 4$ and $\sigma^2 = 400$, respectively.



(a) High SNR

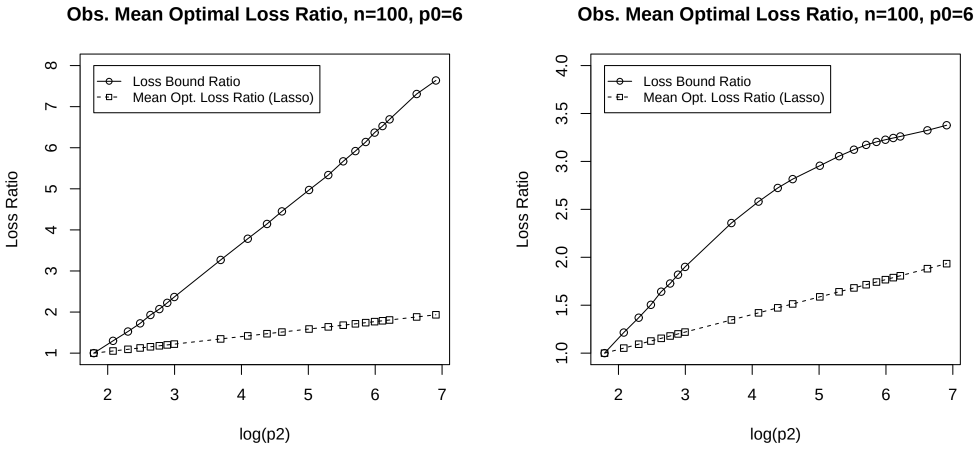
(b) Low SNR

Our simulations suggest that the performance of the Lasso deteriorates for fixed n as p varies. In order to investigate its behavior when n varies, we compare $p_1 = 2\log(n)$ against $p_2 = n$. Under this set-up, p increases as n increases, which is consistent with the standard settings in high-dimensional data analysis. Figure 3 compares the mean optimal loss ratios over 1000 realizations to the optimal loss ratio predicted by the bounds. These plots suggest that the deterioration persists as n increases, and that the bounds under-predict the observed deterioration.

3.2 Independent Predictors

Here we again assume that $\boldsymbol{y}$ is generated from the model given by (3.1) except in this section the columns of $\boldsymbol{X}$ are independent standard normal random variables. This matrix is simulated once and used for all realizations. This allows us to consider situations where $p > n$. We consider both a high and low SNR setting by taking $\sigma^2 = 9$ and $\sigma^2 = 625$, respectively.

Figure 4: Mean optimal loss ratio over 1000 realizations as a function of $\log(p_2)$ for $n = 100$ and $p_1 = 6$. The number of predictor variables p_2 is varied from 6 to 1000. The “High SNR” and “Low SNR” settings correspond to $\sigma^2 = 9$ and $\sigma^2 = 625$, respectively.



(a) High SNR

(b) Low SNR

Figure 4 compares the means of optimal loss ratios over 1000 realizations to the optimal loss ratio predicted by the bound given in equation (3.2). We again set p_1 equal the true set of six predictors and vary p_2 from six to 1000. The bottom line is the optimal loss ratio predicted by the bound, whereas the top line is the observed mean optimal loss ratios. As in the orthonormal design case, this plot shows that the bounds do not adequately measure the deterioration in performance, and that the performance of the Lasso is sensitive to the number of predictor variables.

4. Concluding Remarks

The Lasso allows the fitting of regression models with a large number of predictor variables, but the resulting cost can be much higher than the commonly held views in the literature would suggest. We proved that when tuned optimally for prediction the performance of the Lasso deteriorates as the number of predictor variables increases with non-zero probability under the assumptions of a sparse true model and an orthonormal deterministic design matrix. Our empirical results further suggest that this deterioration persists as the sample size increases.

In light of this deterioration, data analysts should be careful when using the Lasso with high-dimensional data sets. One possible modification to the procedure is to use the Lasso as a subset selector, but not as an estimation procedure. An implementation of this is the extreme version of the Relaxed Lasso (Meinshausen, 2007), which fits least squares regressions to the Lasso selected subsets. Another possible solution is to screen the predictor variables before fitting the Lasso penalized regression. In screening, the typical goal is to reduce from huge scale to something that is $o(n)$ (Fan and Lv, 2008). However, our results suggest that it is not enough to merely reduce the number of predictors, which implies that how to optimally tune the number of screened predictors is an interesting model selection problem for future research.

A. Additional Technical Results

Proof of Lemma 2.2. First note that

$$\sum_{j=2}^{p_1} |z_j|^2 \leq p_1 \max_{2\leq j\leq p_1} |z_j|^2. \tag{A.1}$$

Next, if $z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon]$, then

$$(\beta_1 - z_1)^2 \leq \varepsilon^2, \tag{A.2}$$

and if (2.2) is satisfied, then

$$\varepsilon^2 + p_1 \max_{2\leq j\leq p_1} |z_j|^2 < \beta_1^2. \tag{A.3}$$

Combining (A.1)-(A.3), it follows that

$$L_{p_1}(0) = \frac{1}{n}(\beta_1 - z_1)^2 + \frac{1}{n}\sum_{j=2}^{p_1} |z_j|^2 \leq \frac{1}{n}\varepsilon^2 + \frac{1}{n}p_1 \max_{2\leq j\leq p_1} |z_j|^2 < \frac{1}{n}\beta_1^2.$$

□

Proof of Lemma 2.3. First note that by zeroing out all of the coefficients, we can achieve a loss equal to $\frac{1}{n}\beta_1^2$, so $L_{p_2}(\lambda^*_{p_2}) \leq \frac{1}{n}\beta_1^2$. If (2.3) holds, then

$$\begin{aligned}
|\beta_1| &< \max_{p_1+1\leq j\leq p_2} |z_j| - \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|) \\
&= \max_{p_1+1\leq j\leq p_2} (|z_j| - \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|))_+ \\
&= \max_{p_1+1\leq j\leq p_2} |\hat{\beta}_{j\lambda_{\max}}|
\end{aligned}$$

where $\lambda_{\max} = \max(|\beta_1 - \varepsilon|, |\beta_1 + \varepsilon|)$. If $z_1 \in [\beta_1 - \varepsilon, \beta_1 + \varepsilon]$, then $|z_1| \leq \lambda_{\max}$. Therefore, for any $\lambda < |z_1|$,

$$\beta_1^2 < \max_{p_1+1\leq j\leq p_2} \hat{\beta}^2_{j\lambda_{\max}} \leq \max_{p_1+1\leq j\leq p_2} \hat{\beta}^2_{j\lambda}.$$

From this it follows that, for any $\lambda < |z_1|$,

$$\frac{1}{n}\beta_1^2 < \frac{1}{n}\max_{p_1+1\leq j\leq p_2} \hat{\beta}^2_{j\lambda} < \frac{1}{n}(\beta_1 - \hat{\beta}_{1\lambda})^2 + \frac{1}{n}\sum_{j=2}^{p_2} \hat{\beta}^2_{j\lambda} = L_{p_2}(\lambda). \tag{A.4}$$

Combining (A.4) with the fact that $L_{p_2}(\lambda) \geq \frac{1}{n}\beta_1^2$ for $\lambda \geq |z_1|$, it follows that $L_{p_2}(\lambda^*_{p_2}) = \frac{1}{n}\beta_1^2$. □

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*, pages 267–281.

Bickel, J., Alexandre, and Tsybakov, B. (2010). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*.

Buhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, pages 169–194.

Craven, P. and Wahba, G. (1978). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31(4):377–403.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space with Discussion. *J.R. Statist. Soc. B*, 70(5):849–911.

Flynn, C. J., Hurvich, C. M., and Simonoff, J. S. (2013). Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models. *Journal of the American Statistical Association*, 108(503):1031–1043.

Greenshtein, E. (2006). Best Subset Selection, Persistence in High-Dimensional Statistical Learning and Optimization Under L_1 Constraint. *Annals of Statistics*, 34(5):2367–2386.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Leng, C., Lin, Y., and Wahba, G. (2006). A Note on the Lasso and Related Procedures in Model Selection. *Statistica Sinica*, 16:1273–1284.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, 52(1):374–393.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2009). A unified framework for high-dimensional analysis of \$M\$-estimators with decomposable regularizers. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1348–1356.

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

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

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "Degrees of Freedom" of the Lasso. *The Annals of Statistics*, 35(5):2173–2192.