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

Cheryl J. Flynn,
New York University, New York, USA

Clifford M. Hurvich
New York University, New York, USA

and Jeffrey S. Simonoff
New York University, New York, USA

†
[Received March 2014]

Summary. 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 probability that the best possible predictive performance of the Lasso deteriorates as p increases can be arbitrarily close to one given a sufficiently high signal to noise

†Address for correspondence: Cheryl J. Flynn, Information, Operations, and Management Sciences Department, Stern School of Business, New York University, 44 West 4th St, New York, NY 10012, USA
E-mail: cflynn@stern.nyu.edu
ratio and sufficiently large $p$. We further demonstrate empirically that the deterioration in performance can be far worse than is commonly suggested in the literature and provide a real data example where deterioration is observed.

**Keywords**: 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

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

where $y$ is a $n \times 1$ response vector, $X$ is a $n \times p$ deterministic matrix of predictor variables, $\beta$ 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 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 of the estimated 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

$$y = \mu + \varepsilon$$

(1.1)

where $\mu$ is a $n \times 1$ unknown mean vector and $\varepsilon$ is a $n \times 1$ random noise vector. The $L_2$ loss
is defined as
\[
L(\lambda) = \frac{||\mu - \hat{\mu}_\lambda||^2}{n} = \frac{||\mu - X\hat{\beta}_\lambda||^2}{n}
\]
(1.2)

where \(\hat{\beta}_\lambda\) is the Lasso estimated vector of coefficients for a specific choice of the regularization parameter \(\lambda \in [0, \infty)\) and \(|| \cdot ||^2\) is the squared Euclidean norm. If the true model is included among the candidate models, then \(\mu = X\beta_0\) for some unknown true coefficient vector \(\beta_0\) and the \(L_2\) loss function takes the form
\[
L(\lambda) = \frac{||X(\beta_0 - \hat{\beta}_\lambda)||^2}{n}.
\]

In most modern applications, it is assumed that \(\beta_0\) is sparse and only has \(p_0\) 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 particular deterministic choice of \(\lambda\), these probability bounds are of the form
\[
L(\lambda) \leq k\sigma^2 \log(p) p_0
\]
(1.3)

(p. 102, Buhlmann and van de Geer, 2011). Here \(\sigma^2\) is the true error variance, and \(k\) is a constant that does not depend on \(n\). 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 estimator. 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 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 oracle inequalities and the way that the Lasso is implemented in practice. In practice \(\lambda\) 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 \(\lambda^* = \arg\min_{\lambda} L(\lambda)\) to be the optimal (infeasible) choice of \(\lambda\) that minimizes the loss function over the regularization path. In what follows, we will focus on the loss of the Lasso evaluated at \(\lambda^*\). This selector provides information about the performance of the method in an absolute sense, and it represents the ultimate goal for any model selection procedure designed for prediction. By the definition of \(\lambda^*\), the loss bound (1.3) still applies for \(\lambda = \lambda^*\). Thus, with this choice of selector it is possible to compare the best-case performance of the method against the commonly held views in the literature with a particular focus on the sensitivity of the loss of the Lasso evaluated at \(\lambda^*\) to the number of predictor variables.

The remainder of this paper is organized as follows. Section 2 presents some theoretical results on the behavior of the Lasso based on a data-dependent choice of \(\lambda\) and proves that the best-case predictive performance can deteriorate as the number of predictor variables is increased. In particular, under the assumption of a sparse true model and orthonormal
predictors, we prove that the probability of deterioration can be arbitrarily close to one for a sufficiently high signal to noise ratio and sufficiently large \( p \), and that the expected amount of deterioration is infinite. 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. Section 4 presents an analysis of HIV data using the Lasso and exemplifies the occurrence of deterioration in practice. Finally, Section 5 presents some final remarks and areas for future research. The appendix includes some additional technical and simulation results.

2. Theoretical Results

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

\[
y = X\beta_0 + \varepsilon
\]

where \( y \) is the \( n \times 1 \) response vector, \( X \) is a \( n \times p \) matrix of deterministic predictors such that \( X^TX = I \) (the \( p \times p \) identity matrix), \( \beta_0 = (\beta_1, \ldots, \beta_p)^T \) is the \( p \times 1 \) vector of true unknown coefficients, and \( \varepsilon \) is a \( n \times 1 \) noise vector where \( \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \). We assume that \( \beta_1 \neq 0 \) but all of the other true coefficients are equal to zero. We further assume that there is no intercept. In practice, the intercept is not penalized and is computed as

\[
\bar{y} - \bar{X}\hat{\beta}_\lambda
\]

for each value of \( \lambda \), where \( \bar{y} \) is the mean of the \( y \) and \( \bar{X} \) is a \( 1 \times p \) vector containing the column means of \( X \).

By construction \( z = X^Ty \) is the vector of the least squares-estimated coefficients based on the full model. It follows that

\[
z_1 \sim N(\beta_1, \sigma^2)
\]

(2.1)

and

\[
z_j \overset{iid}{\sim} N(0, \sigma^2)
\]

(2.2)
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 \(\lambda\), 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
\[
L_p(\lambda) = \frac{||X\beta_0 - X\hat{\beta}_\lambda||^2}{n}.
\]
Here we subscript the loss by \(p\) in order to emphasize that the loss at a particular value of \(\lambda\) depends on the number of predictor variables. In particular, for this example,
\[
L_p(\lambda) = \frac{1}{n}(\beta_1 - \hat{\beta}_{\lambda,1})^2 + \frac{1}{n} \sum_{j=2}^{p} \hat{\beta}_{\lambda,j}^2.
\] (2.3)

We wish to study the sensitivity of the Lasso to the number of predictor variables. To do this, we compare the loss when only the one true predictor is used, \(L_1(\lambda)\), to the loss where \(p\) predictors (including the one true predictor) are used, \(L_p(\lambda)\). Under our set-up, if \(p > 1\), then \(L_p(\lambda)/L_1(\lambda) \geq 1\) for any given \(\lambda\). In what follows, we establish the stronger result that \(\min_\lambda L_p(\lambda)/\min_\lambda L_1(\lambda) > 1\) with probability arbitrarily close to one for an appropriately high signal to noise ratio and large \(p\), and that the expected value of this ratio is infinite.

To simplify notation, define
\[
\lambda^*_p = \arg\min_\lambda L_p(\lambda).
\]
Under the orthonormality assumption, we require \(p \leq n\).

**THEOREM 2.1.** For all \(1 < p \leq n\),
\[
\Pr \left( \frac{L_p(\lambda^*_p)}{L_1(\lambda^*_1)} > 1 \right) = \Phi \left( \frac{||\beta_1||}{\sigma} \right) - \frac{1}{2p},
\] (2.4)
where \(\Phi(\cdot)\) is the cumulative distribution function of a standard normal random variable.

To prove Theorem 2.1 we establish the following four lemmas. The proofs are given
On the Sensitivity of the Lasso to the Number of Predictor Variables

First note that one can always choose \( \lambda \geq \max_{1 \leq j \leq p} |z_j| \), which will shrink all of the estimated coefficients to zero. Thus, for all \( p > 0 \), \( nL_p(\lambda^*_p) \leq \beta_1^2 \). The following Lemma establishes that equality always occurs if the sign of \( z_1 \) is incorrect.

**Lemma 2.1.** If \( \text{sgn}(\beta_1) \neq \text{sgn}(z_1) \), then \( nL_p(\lambda^*_p) = \beta_1^2 \) for all \( 0 < p \leq n \).

Lemma 2.1 establishes that if the sign of \( z_1 \) is incorrect, \( L_p(\lambda^*_p) = L_1(\lambda^*_1) \) for all \( p > 1 \), so no deterioration will occur.

Next we focus our attention on the situation where the sign of \( z_1 \) is correct. The following lemma establishes the optimal loss for the Lasso when only the one true predictor is used.

**Lemma 2.2.** If \( \text{sgn}(\beta_1) = \text{sgn}(z_1) \), then

\[
nL_1(\lambda^*_1) = \begin{cases} 
0 & \text{if } |\beta_1| \leq |z_1| \\
(\beta_1 - z_1)^2 & \text{otherwise}
\end{cases}.
\]

In this case, when the model includes superfluous predictors, the optimal level of shrinkage is determined by balancing the increase in loss due to the bias induced from over-shrinking the true estimated coefficient with the increase in loss due to under-shrinking the estimated coefficients for the superfluous predictors. The next two lemmas establish necessary and sufficient conditions on the \( z_j \)'s for deterioration to occur.

**Lemma 2.3.** Assume that \( \text{sgn}(\beta_1) = \text{sgn}(z_1) \). If \( \max_{2 \leq j \leq p} |z_j| < |z_1| \), then \( L_p(\lambda^*_p) = L_1(\lambda^*_1) \) if and only if \( |\beta_1| < |z_1| - \max_{2 \leq j \leq p} |z_j| \).

**Lemma 2.4.** Assume that \( \text{sgn}(\beta_1) = \text{sgn}(z_1) \). If \( \max_{2 \leq j \leq p} |z_j| > |z_1| \), then \( L_p(\lambda^*_p) > L_1(\lambda^*_1) \) for all \( \beta_1 \neq 0 \).

It follows that deterioration occurs unless it is possible to shrink \( z_1 \) optimally while at the same time shrinking all of the estimated coefficients for the superfluous predictors to zero. In particular, by Lemmas 2.3 and 2.4 when the sign of \( z_1 \) is correct, \( L_p(\lambda^*_p) > L_1(\lambda^*_1) \) unless \( |\beta_1| < |z_1| - \max_{2 \leq j \leq p} |z_j| \).
With these results in place, the two terms on the right-hand side of equation (2.4) can be explained intuitively. Specifically, the first term reflects the increasing likelihood that the sign of $z_1$ is correct as the signal-to-noise ratio increases, and the second term reflects the decreasing probability of no deterioration in this case as $p$ increases. We can also establish the probability of deterioration given that the sign of $z_1$ is correct.

**Theorem 2.2.** For all $1 < p \leq n$,

$$
Pr\left( \frac{L_p(\hat{\lambda}_p^*)}{L_1(\hat{\lambda}_1^*)} > 1 \mid \text{sgn}(z_1) = \text{sgn}(\beta_1) \right) = 1 - \frac{1}{2p \Phi\left( \frac{\|\beta_1\|}{\sigma} \right)},
$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

Theorems 2.1 and 2.2 establish that the best case predictive performance of the Lasso can deteriorate as $p$ increases. Under our set-up, the probability of deterioration can be arbitrarily close to one, and the following proposition establishes that the expected amount of deterioration is infinite.

**Proposition 2.5.** For all $1 < p \leq n$,

$$
E\left( \frac{L_p(\hat{\lambda}_p^*)}{L_1(\hat{\lambda}_1^*)} \right) = \infty.
$$

**Example.** To demonstrate the implications of Theorem 2.1, consider an ANOVA model based on an orthonormal regression matrix. Specifically, assume that we have $p$ binary predictor variables, each of which is coded using effects coding, and a balanced design with an equal number of observations falling into each of the $2^p$ combinations. If we scale these predictors to have unit variance, then an ANOVA model on only the main effects is equivalent to a regression on these predictors. Similarly, if we consider all pairwise products and then standardize, a regression including them as well as the main effects is equivalent to an ANOVA with all two-way interactions. We can continue to add higher-order interactions in a similar manner, where a model with all $k$-way interactions includes $\sum_{i=1}^{k} \binom{p}{i}$ predictors.
Table 1. The probability of deterioration when only the main effect of the first predictor has a nonzero effect, $\beta_1 = 3$, $\sigma = 1$, and higher order interactions are included.

| Model                        | Probability of Deterioration |
|------------------------------|-----------------------------|
| Main Effects                 | $p = 2$                     |
|                              | 0.7487                      |
|                              | $p = 4$                     |
|                              | 0.8737                      |
|                              | $p = 6$                     |
|                              | 0.9154                      |
|                              | $p = 8$                     |
|                              | 0.9362                      |
|                              | $p = 10$                    |
|                              | 0.9487                      |
| Two-Way Interactions        | $p = 2$                     |
|                              | 0.8362                      |
|                              | $p = 4$                     |
|                              | 0.9487                      |
|                              | $p = 6$                     |
|                              | 0.9749                      |
|                              | $p = 8$                     |
|                              | 0.9848                      |
|                              | $p = 10$                    |
|                              | 0.9896                      |
| Three-Way Interactions      | $p = 2$                     |
|                              | −                            |
|                              | $p = 4$                     |
|                              | 0.9602                      |
|                              | $p = 6$                     |
|                              | 0.9865                      |
|                              | $p = 8$                     |
|                              | 0.9933                      |
|                              | $p = 10$                    |
|                              | 0.9958                      |
| Four-Way Interactions       | $p = 2$                     |
|                              | −                            |
|                              | $p = 4$                     |
|                              | 0.9630                      |
|                              | $p = 6$                     |
|                              | 0.9898                      |
|                              | $p = 8$                     |
|                              | 0.9956                      |
|                              | $p = 10$                    |
|                              | 0.9974                      |

Assume that only the main effect of the first predictor has a nonzero effect, $\beta_1 = 3$, and that $\sigma = 1$. Then applying the result of Theorem 2.1 Table I shows that the probability of deterioration can be close to one for even a moderate number of predictor variables.

3. Empirical Study

This section empirically investigates the cost of not knowing the true set of predictors when working with high-dimensional data. We assume that $y$ is generated according to the generating model in (1.1). The Lasso regressions are fit using the R `lars` package (Hastie and Efron, 2011). We use the default package settings and include an intercept in the model. 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 $X$ 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 $X$ 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. To measure the deterioration in optimal loss we consider the optimal loss ratio

$$
\frac{L_p(\lambda_p)}{L_{p_0}(\lambda_{p_0})},
$$

which compares the minimum loss based on $p$ predictors to the minimum loss based on the true set of $p_0$ predictors. These $p_0$ predictors have nonzero coefficients. All other coefficients are zero. Here $p_0 < p$ and the $p_0$ true predictors are always a subset of the $p$ predictors. We focus on cases where $p$ is large or is getting large in order to be consistent
3.1. Orthogonal Predictors

Define the true model to be

\[ y_i = 6x_{i,1} + 5x_{i,2} + 4x_{i,3} + 3x_{i,4} + 2x_{i,5} + x_{i,6} + \varepsilon_i \]  

for \( i = 1, \ldots, n \), where \( \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \). 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 \( X \) are trigonometric predictors defined by

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

and

\[ x_{i,2j} = \cos \left( \frac{2\pi j}{n}(i - 1) \right) \]

for \( j = 1, \ldots, p/2 \) and \( i = 1, \ldots, n \). The columns of \( X \) are orthogonal under this design and the true model is always included among the candidate models.
Fig. 1. Median optimal loss ratio over 1000 realizations as a function of \( \log(p) \) for \( n = 100 \) and \( p_0 = 6 \). The number of predictor variables \( p \) is varied from 6 to 100. The “High SNR” and “Low SNR” settings correspond to \( \sigma^2 = 4 \) and \( \sigma^2 = 400 \), respectively.

By the definition of the optimal loss, the oracle inequalities in the literature also apply to \( L_p(\lambda_p^*) \). In what follows, we compare the empirical performance of the optimal loss to two established bounds. First, by applying Corollary 6.2 in Buhlmann and van de Geer (2011),

\[
L_p(\lambda_p^*) \leq 64\sigma^2 p_0 \frac{t^2 + 2\log(p)}{n \psi_0^2} \tag{3.3}
\]

with probability greater than \( 1 - 2e^{-t^2/2} \), where \( \psi_0 \) is a constant that satisfies a compatibility condition. This condition places a restriction on the minimum eigenvalue of \( X^TX/n \) for a restricted set of coefficients and it’s sufficient to take \( \psi_0 = 1 \) for an orthogonal design matrix. Second, by Theorem 6.2 in Bickel et al. (2010),

\[
L_p(\lambda_p^*) \leq 16A^2 \sigma^2 p_0 \frac{\log(p)}{n \kappa^2} \tag{3.4}
\]

with probability at least \( 1 - p^{1-A^2/8} \), where \( \kappa \) is a constant tied to a restricted eigenvalue assumption. For orthogonal predictors, \( \kappa = 1 \). In the simulations, \( t \) and \( A \) are both set so that the bounds hold with at least 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 (roughly) equal to $\log(p)$?

Figure 1 compares the medians of optimal loss ratios over 1000 realizations to the ratio suggested by the loss bound for varying values of $p$. In both plots, the bottom two lines are the loss ratios implied by assuming that the bound equals the optimal loss, whereas the top line is the observed median optimal loss ratio. Comparing the two plots, the deterioration is worse in the High SNR case. This is consistent with our theoretical results, which established that we are more likely to observe deterioration when the SNR is high. When the SNR is low, it is more likely that the optimal loss will equal the loss for $\lambda = \lambda_p^{\text{max}}$, where $\lambda_p^{\text{max}}$ is equal to the value of $\lambda$ that sets all of the $p$ estimated coefficients equal to zero. When this is the case, no further deterioration can occur when adding more superfluous variables.

Clearly the deterioration is far worse than is suggested by the bounds for both choices of the SNR. For example, if we include $n$ predictors in the High SNR case, then the loss bounds suggest we should be about 50 percent worse off than if we knew the true set of predictors, but in actuality we are typically more than 350 percent worse off. This discrepancy is a consequence of the fact that the bounds are inequalities rather than equalities.
Fig. 2. Ratio of the loss bounds to the observed median 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.

To emphasize the danger of relying only on bounds, Figure 2 plots the ratio of the bounds to the median 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 $\log(p)$. 
Fig. 3. Median 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.

The results presented thus far 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 \( p_2 = n \) and define the optimal loss ratio to be

\[
\frac{\min_{\lambda} L_{p_2}(\lambda)}{\min_{\lambda} L_{p_1}(\lambda)}.
\]

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. Since the slopes with respect to \( n \) are higher than the bounds imply, this further suggests that the deterioration gets worse for larger samples.

3.2. Independent Predictors

Here we again assume that \( y \) is generated from the model given by (3.2) except in this section the columns of \( X \) are independent standard normal random variables. This allows
us to consider situations where $p > n$. This matrix is simulated once and used for all realizations. We consider both a high and low SNR setting by taking $\sigma^2 = 9$ and $\sigma^2 = 625$, respectively.

**Fig. 4.** Median optimal loss ratio over 1000 realizations as a function of $\log(p)$ for $n = 100$ and $p_0 = 6$. The number of predictor variables $p$ is varied from 6 to 1000. The “High SNR” and “Low SNR” settings correspond to $\sigma^2 = 9$ and $\sigma^2 = 625$, respectively.

Figure 4 compares the medians of the optimal loss ratios as defined in (3.1) over 1000 realizations to the optimal loss ratio suggested by the bounds (3.3) and (3.4). We vary $p$ from six to 1000. In both plots, the bottom two lines are the optimal loss ratios predicted by the bounds, whereas the top line is the observed median 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 optimal performance of the Lasso is sensitive to the number of predictor variables. Also, we again see that less deterioration occurs in the Low SNR case, where it is more likely that the optimal choice for $\lambda$ is equal to $\lambda_{p}^{\text{max}}$.

4. **Real Data Analysis**

In numerous applications it is desirable to model higher-order interactions; however, the inclusion of such interactions can greatly increase the computational burden of a regression
Table 2. The number of samples and mutation sites in each of the six HIV-1 drug datasets.

| Drug | 3TC | ABC | AZT | D4T | DDI | TDF |
|------|-----|-----|-----|-----|-----|-----|
| n    | 1057| 1005| 1067| 1073| 1073| 784 |
| p    | 217 | 211 | 218 | 218 | 218 | 216 |

analysis. The Lasso provides a computationally feasible solution to this problem.

As an example of this, Bien et al. (2013) recently used the Lasso to investigate the inclusion of all pairwise interactions in the analysis of six HIV-1 drug datasets. The goal of this analysis was to understand the impact of mutation sites on antiretroviral drug resistance. These datasets were originally studied by Rhee et al. (2006) and include a measure of (log) susceptibility for different combinations of mutation sites for each of six nucleoside reverse transcriptase inhibitors. The number of samples (n) and the number of mutation sites (p) for each dataset are listed in Table 2.

In their analysis, Bien et al. (2013) compared the performance of the Lasso with only main effects included in the set of predictors (MEL) to its performance with main effects and all pairwise interactions included (APL). Although not the focus of their analysis, we show here that this application demonstrates the sensitivity of the procedure to the number of predictor variables, which can result in deteriorating performance in the absence of strong interaction effects.

Since the true data-generating mechanism is unknown, we cannot compute the optimal loss ratios for this example. As an alternative, to measure deterioration we randomly split the data into a training- and test-set. We then fit the Lasso using the training-set and evaluate the predictive performance on the test-set by computing the mean square error (MSE). It is important to note that both the numerator and denominator in the MSE ratio include the true error variance, so the loss in estimation precision will be less apparent. To exemplify this, Appendix B studies the optimal MSE ratio in the context of the independent predictors example given in Section 3.2.

Figure 4 plots the ratios of the minimum test-set MSE obtained using the APL to that obtained using the MEL based on 20 random splits of the data for each of the six drugs.
For the ‘3TC’ drug, the inclusion of all pairwise interactions results in a dramatic improvement in performance. In particular, there are five interactions that are included in all twenty of the selected models: ‘p62:p69’, ‘p65:p184’, ‘p67:p184’, ‘p184:p215’, and ‘p184:p190’. This suggests that there is a strong interaction effect in this example, and that the interactions between these molecular targets are useful for the predicting drug susceptibility.

On the other hand, in four of the five remaining drugs - ‘ABC’, ‘D4T’, ‘DDI’, and ‘TDF’ - the inclusion of all pairwise interactions results in a significant deterioration in performance. Here significance is determined using a Wilcoxon signed-rank test performed at a 0.05 significance level. Thus, although the MEL is a restricted version of the APL, we still observe deterioration in the best-case predictive performance. This suggests that although the Lasso allows the modeling of higher-order interactions, their inclusion should
be done with care as doing so can hurt overall performance.

5. Discussion

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 probability arbitrarily close to one 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, and carries over to more general contexts.

In light of this deterioration, data analysts should be careful when using the Lasso as a variable selection and estimation tool 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. Returning to the orthogonal predictors example in Section 3.1, we investigate the performance of this simple two-step procedure. Figure 6 plots the median optimal loss for the Lasso and the median optimal loss for the two-stage procedure for varying values of $p$. In this example, the two-stage procedure improves performance when the SNR is high, but not when the SNR is low. However, the improvement in performance in the high SNR case is more than the worsening of performance in the low SNR case. These preliminary results suggest that a two-stage procedure can help improve performance when the SNR is sufficiently high.

Another possible solution is to screen the predictor variables before fitting the Lasso penalized regression. In screening, the typical goal is to reduce from a 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. Investigating proper screening techniques and two-stage procedures are both interesting areas for future
Fig. 6. Median optimal loss for the Lasso and Lasso+OLS over 1000 realizations as a function of $p$ for the orthogonal predictors example with $n = 100$ and $p_0 = 6$. The “High SNR” and “Low SNR” settings correspond to $\sigma^2 = 4$ and $\sigma^2 = 400$, respectively.

A. Technical Results

**Proof (Lemma 2.1).** If $\text{sgn}(\beta_1) \neq \text{sgn}(z_1)$, then for any $\lambda < \max_{1 \leq j \leq p} |z_j|$, \n
$$nL_p(\lambda) = (\beta_1 - \text{sgn}(z_1)(|z_1| - \lambda)_+)^2 + \sum_{j=2}^{p} (|z_j| - \lambda_+)^2 \geq \beta_1^2 + \sum_{j=2}^{p} (|z_j| - \lambda)_+^2 \geq \beta_1^2.$$ \n
Thus $nL_p(\lambda_p^*) = \beta_1^2$.

**Proof (Lemma 2.2).** Without loss of generality assume that $\beta_1 > 0$, and therefore $z_1 > 0$. Consider \n
$$nL_1(\lambda) = (\beta_1 - (z_1 - \lambda)_+)^2.$$ \n
First consider $\lambda \in I = [0, z_1)$. Since $nL_1(\lambda)$ is a convex function for $\lambda \in I$, the minimum occurs at a place where the derivative is zero or when $\lambda = 0$. Taking the derivative with respect to $\lambda \in I$, \n
$$\frac{\partial nL_1(\lambda)}{\partial \lambda} = 2(\beta_1 - (z_1 - \lambda)).$$
Since the derivative is an increasing function of $\lambda$, a minimum occurs at $\lambda = 0$ if the derivative is non-negative at that point. In other words, a minimum occurs at $\lambda = 0$ if $\beta_1 \geq z_1$. Otherwise, a minimum occurs at a point where the derivative is zero. Thus

$$\arg\min_{\lambda \in I} nL_1(\lambda) = \begin{cases} 
    z_1 - \beta_1 & \text{if } 0 \leq \beta_1 < z_1 \\
    0 & \text{if } z_1 \leq \beta_1
\end{cases},$$

and

$$\min_{\lambda \in I} nL_1(\lambda) = \begin{cases} 
    0 & \text{if } 0 \leq \beta_1 < z_1 \\
    (\beta_1 - z_1)^2 & \text{if } z_1 \leq \beta_1
\end{cases}.$$

Next, for $\lambda \geq z_1$, $nL(\lambda) = \beta_1^2$. Since $\min_{\lambda \in I} nL(\lambda) < \beta_1^2$ for all $\beta_1 > 0$, it follows that

$$nL_1(\lambda^*_1) = \begin{cases} 
    0 & \text{if } 0 \leq \beta_1 < z_1 \\
    (\beta_1 - z_1)^2 & \text{if } z_1 \leq \beta_1
\end{cases}.$$

**Proof (Lemma 2.3).** Without loss of generality assume that $\beta_1 > 0$, and therefore $z_1 > 0$. Also assume that $|z_2| > \ldots > |z_p|$. Consider

$$nL_p(\lambda) = (\beta_1 - (z_1 - \lambda)_+)^2 + \sum_{j=2}^p (|z_j| - \lambda)^2_+.$$

First consider $\lambda \in I = [0, z_1)$. Since $nL_p(\lambda)$ is a continuous differentiable function for $\lambda \in I$, local extrema occur at points where the derivative is zero or at a boundary point. Taking the derivative with respect to $\lambda$,

$$\frac{\partial nL_p(\lambda)}{\partial \lambda} = \begin{cases} 
    2(\beta_1 - (z_1 - \lambda)) & \text{if } |z_2| \leq \lambda < z_1 \\
    2(\beta_1 - (z_1 - \lambda)) - 2\sum_{j=2}^k (|z_j| - \lambda) & \text{if } |z_{k+1}| \leq \lambda < |z_k|, \\
    \text{for } k = 2, \ldots, p - 1 \\
    2(\beta_1 - (z_1 - \lambda)) - 2\sum_{j=2}^p (|z_j| - \lambda) & \text{if } 0 \leq \lambda < |z_p|
\end{cases}.$$

Since the derivative is a strictly increasing function of $\lambda$, a minimum occurs at $\lambda = 0$ if the derivative is non-negative at that point. Hence a minimum occurs at $\lambda = 0$ if $\beta_1 > \sum_{j=1}^p |z_j|$. Otherwise a minimum occurs at a point where the derivative is zero. It follows that
\[ \lambda_1^* \equiv \arg \min_{\lambda \in I} nL_p(\lambda) = \begin{cases} 
 z_1 - \beta_1 & \text{if } 0 < \beta_1 \leq z_1 - |z_2| \\
 \frac{\sum_{j=1}^k |z_j| - \beta_1}{k} & \text{if } \sum_{j=1}^k |z_j| < \beta_1 \leq \sum_{j=1}^k |z_j| - k|z_{k+1}|, \\
 \frac{\sum_{j=1}^p |z_j| - \beta_1}{p} & \text{for } k = 2, \ldots, p - 1 \\
 0 & \text{if } \sum_{j=1}^p |z_j| < \beta_1 
\end{cases}. \]

Next, for \( \lambda \geq z_1 \), \( nL_p(\lambda) = \beta_1^2 \). Thus

\[ nL_p(\lambda_p^*) = \min \left( \beta_1^2, nL_p(\lambda_1^*) \right). \]

To compare \( nL_p(\lambda_p^*) \) to \( nL_1(\lambda_1^*) \), first note that \( nL_1(\lambda_1^*) < \beta_1^2 \). Next, comparing \( nL_p(\lambda_j^*) \) to \( nL_1(\lambda_1^*) \) it is clear that \( nL_1(\lambda_1^*) = nL_p(\lambda_j^*) = 0 \) if \( 0 < \beta_1 \leq z_1 - |z_2| \). However, if \( z_1 - |z_2| < \beta_1 \leq z_1 \), then \( \lambda_j^* < |z_2| \) and

\[ nL_p(\lambda_j^*) > (|z_2| - \lambda_p^*)^2 > 0 = nL_1(\lambda_1^*). \]

Similarly, if \( z_1 < \beta_1 \), then either \( \lambda_j^* > 0 \) so that

\[ nL_p(\lambda_j^*) > (\beta_1 - (z_1 - \lambda_j^*))^2 > (\beta_1 - z_1)^2 = nL_1(\lambda_1^*), \]

or \( \lambda_j^* = 0 \) and

\[ nL_p(\lambda_j^*) = (\beta_1 - z_1)^2 + \sum_{j=2}^p |z_j|^2 > (\beta_1 - z_1)^2 = nL_1(\lambda_1^*). \]

Hence, \( nL_1(\lambda_1^*) = nL_p(\lambda_p^*) \) if and only if \( 0 < \beta_1 \leq z_1 - |z_2| \).

PROOF (LEMMA 2.4). Without loss of generality assume that \( \beta_1 > 0 \), and therefore \( z_1 > 0 \). Also assume that \( |z_2| > \ldots > |z_p| \). Consider

\[ nL_p(\lambda) = (\beta_1 - (z_1 - \lambda))^2 + \sum_{j=2}^p (|z_j| - \lambda_j^*)^2. \]
Define
\[ \tilde{k} = \max_{2 < k \leq p} \{ k : |z_k| > z_1 \} . \]
The derivative of \( nL_p(\lambda) \) does not exist at \( \lambda = z_1 \). However, by Lemma 2.1, \( \lambda = z_1 \) is never globally optimal since
\[ nL_p(z_1) = \beta_1^2 + \sum_{j=2}^{\tilde{k}} (|z_j| - z_1)^2 > \beta_1^2. \]
To determine the optimal values of \( \lambda \), we consider the intervals \( I_1 = [0, z_1) \), \( I_2 = (z_1, |z_2|] \), and \( I_3 = (|z_2|, \infty) \) separately. Define
\[ \lambda_{I_j}^* = \arg\min_{\lambda \in I_j} nL_p(\lambda) \]
for \( j = 1, 2, 3 \).
First, for \( \lambda \in I_1 \), \( nL_p(\lambda) \) is a continuous differentiable function and
\[ \frac{\partial nL_p(\lambda)}{\partial \lambda} = 2(\beta_1 - (z_1 - \lambda)) - 2 \sum_{j=2}^{p} (|z_j| - \lambda)_+. \]
Since the derivative is a strictly increasing function of \( \lambda \), a minimum occurs at \( \lambda = 0 \) if the derivative is non-negative at that point. Thus, a minimum occurs at \( \lambda = 0 \) if \( \beta_1 > \sum_{j=1}^{p} |z_j| \).
Similarly, a local minimum occurs at \( \lambda = z_1 \) if
\[ \lim_{\lambda \to z_1} \frac{\partial nL_p(\lambda)}{\partial \lambda} < 0, \]
which holds if \( 0 < \beta_1 \leq \sum_{j=1}^{\tilde{k}} |z_j| - \tilde{k}z_1 \). Otherwise, a minimum occurs at a point where the
derivative is zero. It follows that

\[
\lambda_{I_2}^* = \begin{cases} 
  z_1 & \text{if } 0 < \beta_1 \leq \sum_{j=1}^k |z_j| - \tilde{k}z_1 \\
  \frac{\sum_{j=1}^k |z_j|-\beta_1}{k} & \text{if } \sum_{j=1}^k |z_j| - \tilde{k}z_1 < \beta_1 \leq \sum_{j=1}^k |z_j| - \tilde{k}z_{k+1}, \\
  \frac{\sum_{j=1}^k |z_j|-\beta_1}{k} & \text{if } \sum_{j=1}^k |z_j| - k|z_k| < \beta_1 \leq \sum_{j=1}^k |z_j| - k|z_{k+1}|, \\
  \frac{\sum_{j=1}^p |z_j|-\beta_1}{p} & \text{if } \sum_{j=1}^p |z_j| - p|z_p| < \beta_1 \leq \sum_{j=1}^p |z_j|, \\
  0 & \text{if } \sum_{j=1}^p |z_j| < \beta_1
\end{cases}
\]

Next, for \( \lambda \in I_2 \), \( nL_p(\lambda) \) is a continuous differentiable function and

\[
\frac{\partial nL_p(\lambda)}{\partial \lambda} = -2 \sum_{j=2}^p (|z_j| - \lambda)_+.
\]

Since the derivative is negative for all \( \lambda \in I_2 \), a local minimum occurs at \( \lambda = |z_2| \), thus \( nL_p(\lambda_{I_2}^*) = \beta_2^2 \).

Lastly, for all \( \lambda \in I_3 \), \( nL_p(\lambda) = \beta_1^2 \). It follows that

\[
nL_p(\lambda_{p}^*) = \min \left( \beta_1^2, nL_p(\lambda_{I_1}^*) \right)
\]

By a similar argument to that used in the proof of Lemma \ref{Lemma2.3}, it follows that \( nL_p(\lambda_{p}^*) > nL_1(\lambda_1^*) \).

**Proof (Theorem 2.1).** By Lemma \ref{Lemma2.1}

\[
\Pr \left( nL_p(\lambda_{p}^*) > nL_1(\lambda_1^*) \right) = \Pr \left( nL_p(\lambda_{p}^*) > nL_1(\lambda_1^*) \mid \text{sgn}(z_1) = \text{sgn}(\beta_1) \right) \Pr \left( \text{sgn}(z_1) = \text{sgn}(\beta_1) \right).
\]

Without loss of generality, assume that \( \beta_1 > 0 \). By Lemmas \ref{Lemma2.3} \ref{Lemma2.4}, this is equal to

\[
\left( 1 - \Pr \left( nL_p(\lambda_{p}^*) = nL_1(\lambda_1^*) \mid z_1 > 0 \right) \right) \Pr (z_1 > 0)
\]

\[
= \left( 1 - \Pr \left( \beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| |z_1 > 0 \right) \right) \Pr (z_1 > 0).
\]
We can evaluate these probabilities explicitly. First consider

$$
\Pr(z_1 > 0) = \Phi\left(\frac{\beta_1}{\sigma}\right). \tag{A.1}
$$

Next,

$$
\Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| | z_1 > 0\right) = \frac{\Pr\left(\{\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| \} \cap \{z_1 > 0\}\right)}{\Pr(z_1 > 0)} = \frac{\Pr(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j|)}{\Pr(z_1 > 0)},
$$

where the second equality follows from the fact that $\beta_1 > 0$ implies that $z_1 > 0$. By (2.1) and (2.2),

$$
\Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j|\right) = \Pr\left(\bigcap_{j=2}^p \{\beta_1 < z_1 - |z_j|\}\right) = \frac{1}{2} \int_{-\infty}^{\beta_1} \Phi\left(\frac{z_1 - \beta_1}{\sigma}\right) - 1 d\beta_1 = \frac{1}{2} \int_{-\infty}^{\beta_1} \Phi\left(\frac{z_1 - \beta_1}{\sigma}\right) - 1 d\beta_1,
$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the probability distribution functions (pdf) of $z_1$ and $z_2$, respectively, and $\phi(\cdot)$ is the pdf of the standard normal distribution. Substituting $w = 2\Phi\left(\frac{z_1 - \beta_1}{\sigma}\right) - 1$,

$$
\Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j|\right) = \frac{1}{2} \int_{0}^{1} w^{p-1} dw = \frac{1}{2p}.
$$

Thus

$$
\Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| | z_1 > 0\right) = \frac{1}{2p} \Phi\left(\frac{\beta_1}{\sigma}\right). \tag{A.2}
$$

From (A.1) and (A.2), it follows that

$$
\Pr(L_p(\lambda^*_p) > L_1(\lambda^*_1)) = \Phi\left(\frac{\beta_1}{\sigma}\right) - \frac{1}{2p}.
$$
**Proof (Theorem 2.2).** Consider

\[
\Pr\left(nL_p(\lambda_p^*) > nL_1(\lambda_1^*) \mid \text{sgn}(z_1) = \text{sgn}(\beta_1)\right).
\]

Without loss of generality, assume that \(\beta_1 > 0\). By Lemmas 2.3-2.4,

\[
\Pr\left(nL_p(\lambda_p^*) > nL_1(\lambda_1^*) \mid z_1 > 0\right) = \left(1 - \Pr\left(nL_p(\lambda_p^*) = nL_1(\lambda_1^*) \mid z_1 > 0\right)\right) = \left(1 - \Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| \mid z_1 > 0\right)\right).
\]

From the proof of Theorem 2.1

\[
\Pr\left(\beta_1 < z_1 - \max_{2 \leq j \leq p} |z_j| \mid z_1 > 0\right) = \frac{1}{\frac{2p}{\sigma}} \Phi\left(\frac{\beta_1}{\sigma}\right).
\]

**Proof (Proposition 2.5).** Without loss of generality, assume that \(\beta_1 > 0\). Define

\[\mathcal{A} : = \{z_1, z_2 : z_2 > z_1 > \beta_1\}.
\]

Note that \(\Pr\left((z_1, z_2) \in \mathcal{A}\right) > 0\). By Lemmas 2.2 and 2.4 for \((z_1, z_2) \in \mathcal{A}, L_1(\lambda_1^*) = 0\) and \(L_p(\lambda_p^*) > L_1(\lambda_1^*)\). Thus,

\[
\frac{L_p(\lambda_p^*)}{L_1(\lambda_1^*)} = \infty
\]

It follows that

\[
\mathbb{E}\left(\frac{L_p(\lambda_p^*)}{L_1(\lambda_1^*)}\right) = \int_{z_p = -\infty}^{z_p = \infty} \cdots \int_{z_1 = -\infty}^{z_1 = \infty} \frac{L_p(\lambda_p^*)}{L_1(\lambda_1^*)} df_1(z_1) \cdots df_p(z_p)
\]

\[
\geq \int_{z_p = -\infty}^{z_p = \infty} \cdots \int_{(z_1, z_2) \in \mathcal{A}} \frac{L_p(\lambda_p^*)}{L_1(\lambda_1^*)} df_1(z_1) \cdots df_p(z_p)
\]

\[= \infty.
\]
B. Optimal MSE Ratio

Here we return to the independent predictors example in section 3.2. To study the behavior of the optimal MSE, we evaluate the MSE for each realization on a simulated test set. Figure 7 presents boxplots of the ratios of the estimated optimal MSE with $p$ predictors to the estimated optimal MSE with the six true predictors where $p$ is taken to be 100, 500, and 1000. A comparison of this figure to the median optimal loss ratios presented in Figure 4 demonstrates that while deterioration is still observed, the optimal MSE ratios can be smaller than the optimal loss ratios due to the presence of the error variance in both the numerator and denominator of the MSE ratio.

Fig. 7. Optimal MSE ratio for $p$ predictors compared to the 6 true predictors over 1000 realizations as a function of $p$. The “High SNR” and “Low SNR” settings correspond to $\sigma^2 = 9$ and $\sigma^2 = 625$, respectively.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*, pp. 267–281.
Bickel, J., Alexandre, and B. Tsybakov (2010). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* 37, 1705–1732.

Bien, J., J. Taylor, and R. Tibshirani (2013). A lasso for hierarchical interactions. *Annals of Statistics* 41, 1111–1141.

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

Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* 1, 169–194.

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

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

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

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

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

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

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

Hastie, T. and B. Efron (2011). *lars: Least Angle Regression, Lasso and Forward Stage-wise*. R package version 0.9-8.
C. J. Flynn, C. M. Hurvich and J. S. Simonoff

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

Leng, C., Y. Lin, and G. Wahba (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, 374–393.

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

Rhee, S.-Y., J. Taylor, G. Wadhera, A. Ben-Hur, and D. L. Brutlag (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proceedings of the National Academy of Sciences USA 103, 17355–17360.

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

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

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