Statistical Science
2017, Vol. 32, No. 1, 88–105
DOI: 10.1214/16-STS586

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

Cheryl J. Flynn, Clifford M. Hurvich and Jeffrey S. Simonoff

Abstract. The Lasso is a computationally efficient regression regularization 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 is positive and can be arbitrarily close to one given a sufficiently high signal to noise ratio and sufficiently large p. We further demonstrate empirically that the amount of deterioration in performance can be far worse than the oracle inequalities suggest and provide a real data example where deterioration is observed.

Key words and phrases: 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

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\,\mathrm{Pen}(\boldsymbol{\beta}), \tag{1.1}$$

where $\mathbf{y}$ is a $n \times 1$ response vector, $\mathbf{X}$ is a $n \times p$ deterministic matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients and $\mathrm{Pen}(\cdot)$ 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 area of research.

This paper focuses on the predictive performance of the Lasso and the impact of regularization. To that end, we evaluate the Lasso-estimated models using the l_2-loss function. We assume that the true data generating process is

$$\mathbf{y}=\boldsymbol{\mu}+\boldsymbol{\varepsilon}, \tag{1.2}$$

Cheryl J. Flynn is Senior Inventive Scientist in the Statistics Department in the Big Data Research Organization at AT&T Labs, New York, New York 10007, USA (e-mail: cflynn@research.att.com). Clifford M. Hurvich is Professor of Statistics in the Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University, New York, New York 10012, USA. Jeffrey S. Simonoff is Professor of Statistics in the Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University, New York, New York 10012, USA.

where $\boldsymbol{\mu}$ is a $n \times 1$ unknown mean vector and $\boldsymbol{\varepsilon}$ is a $n \times 1$ random noise vector. Then the l_2-loss is defined as

$$L_p(\lambda) = \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n} = \frac{\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2}{n}, \tag{1.3}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ 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. Here, we subscript the loss by p to emphasize that the loss at a particular value of λ depends on the number of predictor variables. If the true model is included among the candidate models, then $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0$ for some unknown true coefficient vector $\boldsymbol{\beta}_0$ and the l_2-loss function takes the form

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

To be consistent with most modern applications, we allow $\boldsymbol{\beta}_0$ to be sparse and assume that it has $p_0 \leq p$ nonzero entries.

Probability loss bounds exist for the Lasso in this setting (e.g., Candès and Plan, 2009, Bickel, Ritov and Tsybakov, 2009 and Bühlmann and van de Geer, 2011). Roughly, for a particular deterministic choice, λ^0, of λ, these probability bounds are of the form

$$L_p(\lambda^0) \leq k\sigma^2 \frac{\log(p) p_0}{n} \tag{1.4}$$

(Bühlmann and van de Geer, 2011, page 102). Here, σ^2 is the true error variance, and k is a constant that does not depend on n or p. These bounds are commonly termed "oracle inequalities" since, apart from the $\log(p)$ term and the constant, they equal 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 "$\log(p)$-factor is the price to pay by not knowing the active set" (Bühlmann, 2013) and "it is also known that one cannot, in general, hope for a better result" (Candès and Plan, 2009). Under certain assumptions and an appropriate control of the number of predictor variables, these bounds establish l_2-loss consistency in the sense that the l_2-loss will tend to zero asymptotically. Similar upper bounds exist for the expected value of the loss (Bunea, Tsybakov and Wegkamp, 2007a) as well as lower bounds when $\mathbf{X}$ is nonsingular (Chatterjee, 2014). Bunea, Tsybakov and Wegkamp (2006) and Bunea, Tsybakov and Wegkamp (2007b) further established bounds on the loss for random designs and Thrampoulidis, Panahi and Hassibi (2015) studied the asymptotic behavior of the normalized squared error of the Lasso when $p \to \infty$ and $\sigma \to 0$ under the assumption of a Gaussian design matrix. In related work on predictive performance, Greenshtein and Ritov (2004) and Greenshtein (2006) also studied the "persistence" of the Lasso estimator and showed that the difference between the expected prediction error of the Lasso estimator at a particular deterministic value of λ and the optimal estimator converges to zero in probability. Thus, the "Lasso achieves a squared error that is not far from what could be achieved if the true sparsity pattern were known" (Vidaurre, Bielza and Larrañaga, 2013).

Unfortunately, there is a disconnect between these theoretical results 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 by using (k-fold) cross-validation (CV) (see, e.g., Fan and Li, 2001, Leng, Lin and Wahba, 2006, Zou, Hastie and Tibshirani, 2007, Yu and Feng, 2014, Flynn, Hurvich and Simonoff, 2013 and Homrighausen and McDonald, 2014). Since the existing theoretical results do not apply to a data-dependent choice of λ (Chatterjee, 2014), it is not clear how well the oracle inequalities represent the performance of the Lasso in practice.

This motivates us to study the behavior of the loss at a data-dependent choice of the regularization parameter. We define the random variable $\lambda_p^* = \operatorname{argmin}_\lambda L_p(\lambda)$ to be the optimal (infeasible) choice of λ that minimizes the loss function over the regularization path. In what follows, we focus on the loss of the Lasso evaluated at λ_p^*. 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 the optimal loss, the oracle inequalities in the literature also apply to $L_p(\lambda_p^*)$. It is therefore tempting to use the oracle inequalities in the literature to describe the behavior of the optimal loss. The work on persistency has also led to conclusions such as "there is 'asymptotically no harm' in introducing many more explanatory variables than observations" (Greenshtein and Ritov, 2004), and that "in some 'asymptotic sense', when assuming a sparsity condition, there is no loss in letting [p] be much larger than n" (Greenshtein, 2006). More generally, when working in high-dimensional settings these results are interpreted to imply that "having too many components does not degrade forecast accuracy" (Hyndman, Booth and Yasmeen, 2013) and "it will not hurt to include more variables" (Lin, Foster and Ungar, 2011). However, it is important to remember that the existing theoretical results are based on inequalities, not equalities, so they do not necessarily describe the behavior of the optimal loss or the cost of working in high-dimensional settings. To our knowledge, this is the first explicit study of the sensitivity of the best-case predictive performance 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 λ and proves that the best-case predictive performance can deteriorate as the number of predictor variables is increased, in the sense that best-case performance worsens as superfluous variables are added to the set of predictors. In particular, under the assumption of a sparse true model and orthonormal predictors, we prove that the probability of deterioration is nonzero. In the special case where there is only one true predictor, we further 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 amount of deterioration empirically and shows that it can be much worse than one might expect from looking at the loss bounds in the literature. 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

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

where $\mathbf{y}$ is the $n \times 1$ response vector, $\mathbf{X}$ is a $n \times p$ matrix of deterministic predictors such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ (the $p \times p$ identity matrix), $\boldsymbol{\beta}_0 = (\beta_1, \dots, \beta_p)^T$ is the $p \times 1$ vector of true unknown coefficients and $\boldsymbol{\varepsilon}$ is a $n \times 1$ noise vector where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Under the orthonormality assumption, we require $p \le n$.

We define p_0 to be the number of nonzero true coefficients, where $1 \le p_0 \le p$. Without loss of generality, we assume that $\beta_j \neq 0$ for $1 \le j \le p_0$ and $\beta_j = 0$ for $p_0 < j \le p$. We further assume that there is no intercept.

By construction, $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ is the vector of the least squares-estimated coefficients based on the full model. It follows that the z_j's are independent for all $1 \le j \le p$, and that

$$z_j \sim N(\beta_j, \sigma^2) \tag{2.1}$$

for $1 \le j \le p_0$ and

$$z_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{2.2}$$

for $p_0 < j \le p$. For a given λ, the Lasso estimated coefficients are

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

for $j = 1, \dots, p$ (Fan and Li, 2001). We use $L_p(\lambda)$ to measure the performance of this estimator. Under our set-up,

$$L_p(\lambda) = \frac{1}{n}\sum_{j=1}^{p_0}(\beta_j - \hat{\beta}_{\lambda j})^2 + \frac{1}{n}\sum_{j=p_0+1}^{p} \hat{\beta}_{\lambda j}^2. \tag{2.3}$$

We wish to study the sensitivity of the Lasso to the number of predictor variables and to investigate the occurrence of deterioration in practice. Recall that deterioration is defined to be the worsening of best-case performance as superfluous variables are added to the set of predictors. Thus, deterioration occurs when the optimal loss ratio

$$\frac{L_p(\lambda_p^*)}{L_{p_0}(\lambda_{p_0}^*)} > 1$$

for $p > p_0$.

In what follows, we establish that the best case predictive performance of the Lasso deteriorates as p increases with nonzero probability. For ease of presentation, the proofs for the technical results in this section are presented in Appendix A.

THEOREM 2.1. *For all* $1 \le p_0 < p \le n$,

$$\Pr\left(\frac{L_p(\lambda_p^*)}{L_{p_0}(\lambda_{p_0}^*)} > 1\right) > 0. \tag{2.4}$$

To prove Theorem 2.1, we make use of the following lemma, which establishes the conditions under which deterioration occurs.

LEMMA 2.1. *For all* $1 \le p_0 < p \le n$,

$$\frac{L_p(\lambda_p^*)}{L_{p_0}(\lambda_{p_0}^*)} > 1$$

if and only if

$$\lambda^*_{p_0} < \max_{1 \le j \le p_0} |z_j|$$

and

$$\lambda^*_{p_0} < \max_{p_0 < j \le p} |z_j|.$$

To understand the results of Lemma 2.1, first note that for all $p > 0$, $L_p(\lambda^*_p) \le \frac{1}{n}\sum_{j=1}^{p_0} \beta_j^2$, because there always exists a λ such that all of the estimated coefficients are shrunk to zero. Thus, no deterioration occurs in the extreme case where $\lambda^*_{p_0}$ is equal to such a value. In particular, this occurs if $\lambda^*_{p_0} \ge \max_{1 \le j \le p_0} |z_j|$. Outside of this case, the optimal loss will deteriorate if we cannot set the estimated coefficients for the extraneous predictors equal to zero without imposing more shrinkage on the estimated coefficients for the true predictors. This occurs if $\lambda^*_{p_0} \ge \max_{p_0 < j \le p} |z_j|$.

In the special case where $p_0 = 1$, it is further possible to derive a simple exact expression for the probability of deterioration.

THEOREM 2.2. *For $p_0 = 1$ and for all $1 < p \le n$,*

$$\mathrm{Pr}\left(\frac{L_p(\lambda^*_p)}{L_{p_0}(\lambda^*_{p_0})} > 1\right) = \Phi\left(\frac{|\beta_1|}{\sigma}\right) - \frac{1}{2p}, \tag{2.5}$$

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

In Appendix A, we establish that when $p_0 = 1$ $nL_p(\lambda^*_p) = \beta_1^2$ for all $p > 0$ if the sign of z_1 is incorrect. This means that no deterioration occurs in this case. With this result in place, the two terms on the right-hand side of equation (2.5) can be explained intuitively. 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. This result establishes that deterioration occurs with probability arbitrarily close to one for an appropriately high signal to noise ratio and large p when $p_0 = 1$, and the following theorem establishes that the expected amount of deterioration is infinite.

THEOREM 2.3. *For $p_0 = 1$ and for all $1 < p \le n$,*

$$\mathrm{E}\left(\frac{L_p(\lambda^*_p)}{L_{p_0}(\lambda^*_{p_0})}\right) = \infty.$$

The result of Theorem 2.3 follows from the fact that the case where $L_{p_0}(\lambda^*_{p_0}) = 0$ and $L_p(\lambda^*_p) > 0$ occurs with nonzero probability when $p_0 = 1$. We further investigate the amount of deterioration in the more general p_0-sparse case using simulations in Section 3.

As an alternative to loss, performance could also be measured based on Mean Squared Error (MSE). Under the assumption of a deterministic design matrix,

$$\begin{aligned}\mathrm{MSE}_p(\lambda) = E^*\left(\frac{\|\mathbf{y}^* - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n}\right) &= \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n} + \frac{\sigma^2}{n} \\ &= L_p(\lambda) + \frac{\sigma^2}{n},\end{aligned}$$

where $\mathbf{y}^*$ is from an independent test set and the expectation E^* is taken with respect to this independent test set. Thus, Theorems 2.1–2.2 also apply to MSE. Since MSE also includes the error variance, the relative deterioration of MSE is expected to be less than that of loss when using the one correct predictor. We discuss this further in our real data application in Section 4 where we study deterioration in average squared prediction error.

EXAMPLE. To demonstrate the implications of Theorem 2.2, 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.

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 2.2, Table 1 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 $\mathbf{y}$ is generated by the model in (1.2). The Lasso regressions are fit using the R `glmnet` package (Friedman, Hastie and Tibshirani, 2010). We use the default package settings and include an intercept in the model. We consider two simulation set-ups. The first studies the performance of the

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

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

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

$$\frac{L_p(\lambda_p^*)}{L_{p_0}(\lambda_{p_0}^*)}, \tag{3.1}$$

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 grows with n in order to be consistent with high-dimensional frameworks.

By the definition of λ_p^*, 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 (computed over the default grid of λ values) to two established bounds. First, by applying Corollary 6.2 in Bühlmann and van de Geer (2011),

$$L_p(\lambda_p^*) \le 64\sigma^2 p_0 \frac{t^2 + 2\log(p)}{n\psi_0^2} \tag{3.2}$$

with probability greater than $1 - 2e^{-t^2/2}$ for any constant $t > 0$, where ψ_0 is a constant that satisfies a compatibility condition. This condition places a restriction on the minimum eigenvalue of $\mathbf{X}^T\mathbf{X}/n$ for a restricted set of coefficients and it is sufficient to take $\psi_0 = 1$ for an orthogonal design matrix. Second, by Theorem 6.2 in Bickel, Ritov and Tsybakov (2009),

$$L_p(\lambda_p^*) \le 16A^2\sigma^2 p_0 \frac{\log(p)}{n\kappa^2} \tag{3.3}$$

with probability at least $1 - p^{1-A^2/8}$ for any constant $A > 0$, where κ 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 by comparing the observed optimal loss ratio to the loss bound ratio. Here, we define the loss bound ratio to be the ratio that compares each bound based on p predictors to the corresponding bound based on p_0 predictors. The results based on (3.2) and (3.3) are similar in the simulation examples in Sections 3.1 and 3.2, so only the results for (3.2) are reported.

In addition to the infeasible optimal loss, we also consider the performance of the Lasso when tuned using 10-fold CV. For each simulation, we denote the CV-selected λ by λ_p^{CV} with corresponding loss $L_p(\lambda_p^{\mathrm{CV}})$. The CV loss ratio is then computed as

$$\frac{L_p(\lambda_p^{\mathrm{CV}})}{L_{p_0}(\lambda_{p_0}^{\mathrm{CV}})}.$$

Although the bounds in equations (3.2) and (3.3) are not guaranteed to hold for λ_p^{CV}, we compare the observed CV loss ratios to the loss bound ratios to determine how well they predict the Lasso's performance in practice.

3.1 Orthogonal Predictors

Define the true model to be

$$\begin{aligned} y_i &= 6x_{i,1} + 5x_{i,2} + 4x_{i,3} + 3x_{i,4} + 2x_{i,5} \\ &\quad + x_{i,6} + \varepsilon_i \end{aligned} \tag{3.4}$$

for $i = 1, \dots, n$, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. 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

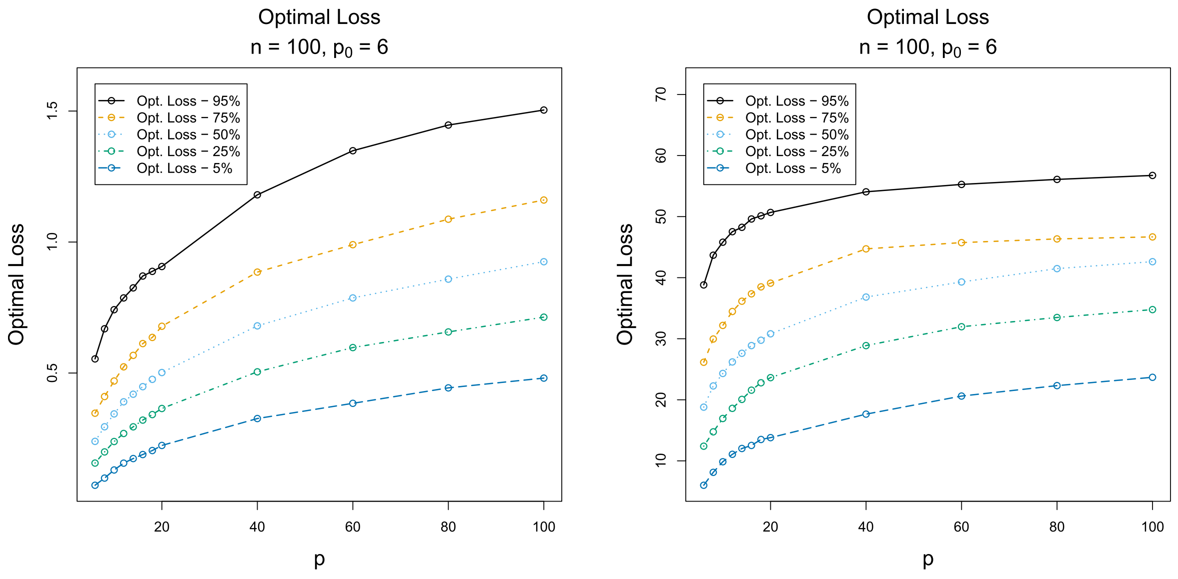
$$x_{i,2j-1} = \sin\left(\frac{2\pi j}{n}(i-1)\right)$$

and

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

for $j = 1, \dots, p/2$ and $i = 1, \dots, n$. The columns of **X** are orthogonal under this design and the true model is always included among the candidate models.

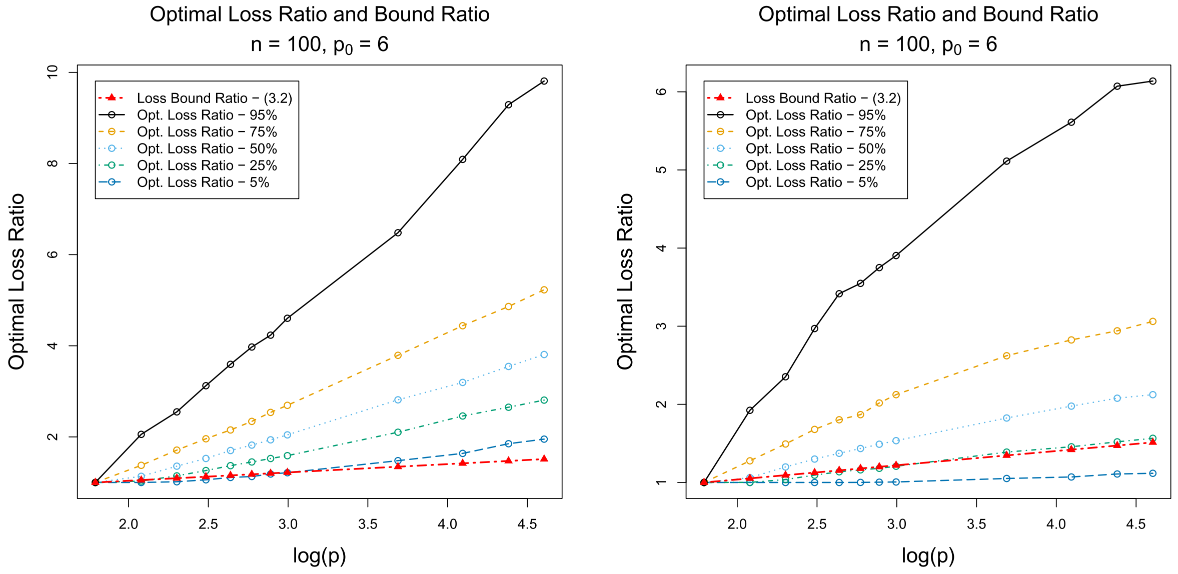
We first compute the optimal loss, $L_p(\lambda_p^*)$, for varying values of p over 1000 realizations. Figure 1 plots



(a) High SNR (b) Low SNR

FIG. 1. *Optimal loss percentiles over* 1000 *realizations as a function of* 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.*

the percentiles of the optimal losses as a function of p. In both the high and low SNR settings, there are signs of deterioration in optimal performance as the number of predictor variables increases, as evidenced by the positive slopes of the percentiles as p increases. To compare this deterioration to the bounds, Figure 2 plots the percentiles of the optimal loss ratios over 1000 realizations and the ratio suggested by the loss bound for varying values of p. In both plots, the loss ratios implied by assuming that the bound equals the optimal loss typically under-estimate the observed 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



(a) High SNR (b) Low SNR

FIG. 2. *Percentiles of the optimal loss ratios 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.*

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 $\lambda = \lambda_p^{\max}$, where $\lambda_p^{\max}$ is equal to the value of λ 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 amount of deterioration is typically far worse than is suggested by the bounds for both choices of the SNR. For example, looking at the median optimal loss ratio, if we include $n = 100$ 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 300 percent worse off. This discrepancy is a consequence of the fact that the bounds are inequalities rather than equalities.

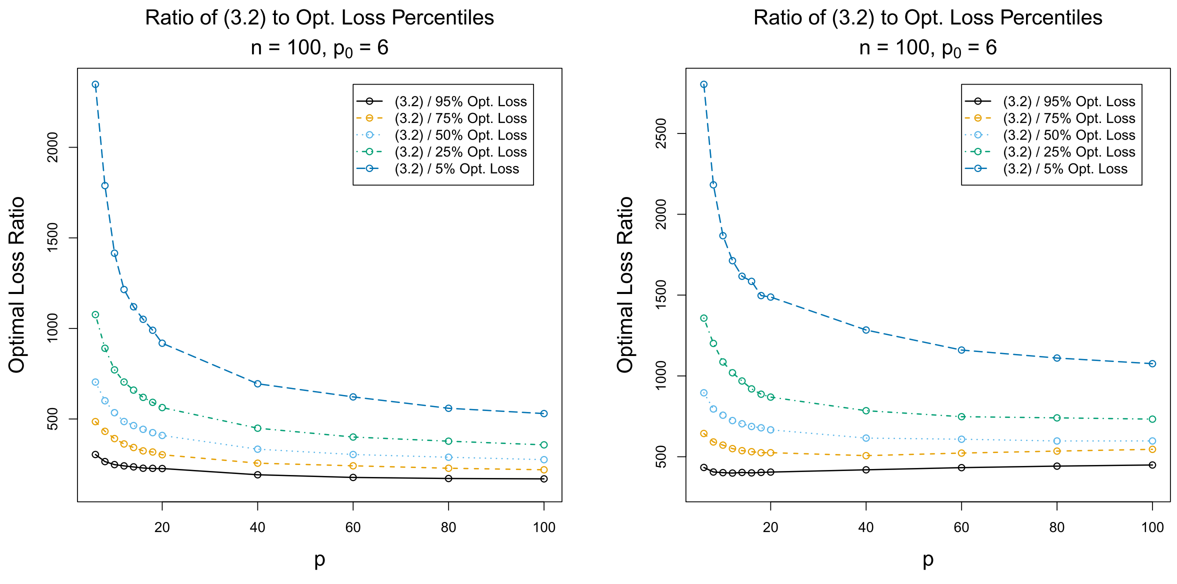
To emphasize the danger of over-interpreting the bounds, Figure 3 plots the ratio of the bounds to the optimal loss percentiles for varying values of p. These plots suggest that the bounds are overly conservative when compared to the optimal loss and the degree of conservatism depends on both p and the SNR. Thus, although the bounds apply, the slope of the optimal loss as a function of p is different than the slope suggested by the bound. As a result of this behavior, the amount of deterioration in optimal loss can be much worse than the bounds suggest. To provide further insight, Figure 4 plots the average ratio of λ^0 to λ_p^* plotted on a log-scale [recall that λ_0 is the deterministic choice of λ used in the oracle inequality (1.4)]. These plots indicate that λ_p^* is typically much smaller than λ^0.

The optimal selector provides the best-case performance of the Lasso, but it is infeasible in practice. This motivates us to also study the performance of the Lasso when λ is selected in a feasible manner using 10-fold CV. Figure 5 compares the CV loss ratios to the bound ratios for varying values of p in the high and low SNR settings. Similar to the optimal loss, we observe deterioration in the CV loss as p increases that is typically worse than the deterioration suggested by the bounds in both SNR settings.

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

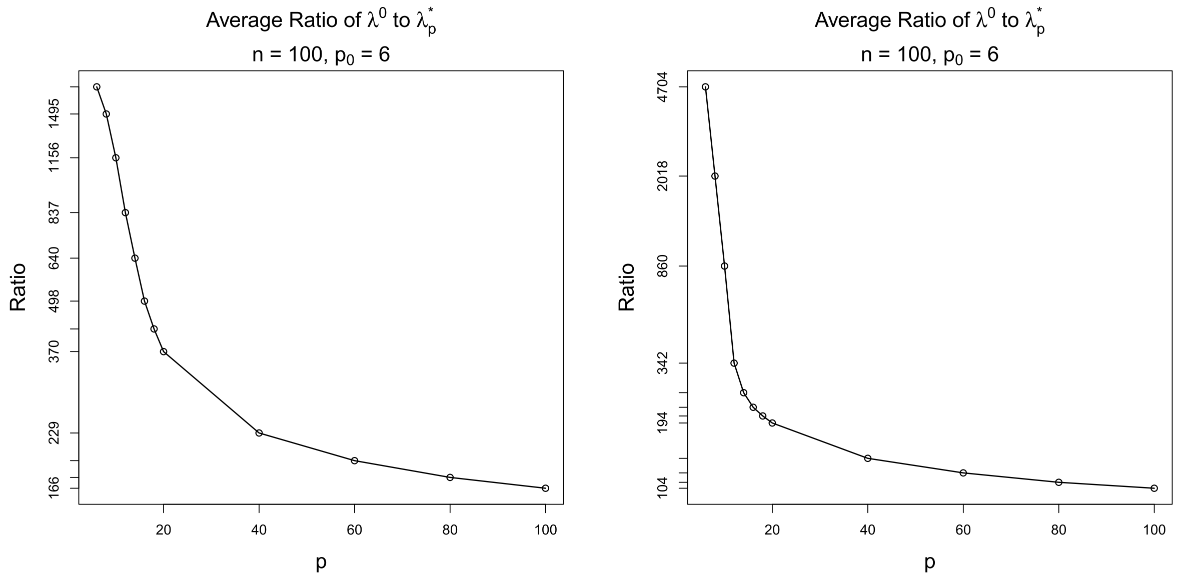
$$\frac{L_{p_2}(\lambda_{p_2}^*)}{L_{p_1}(\lambda_{p_1}^*)}.$$

Under this set-up, p increases as n increases, which is consistent with the standard settings in high-dimensional data analysis. Figure 6 compares the percentiles of the optimal loss ratios over 1000 realizations to the optimal loss ratio suggested 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



(a) High SNR (b) Low SNR

FIG. 3. *Ratio of the loss bounds to the observed optimal loss percentiles 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

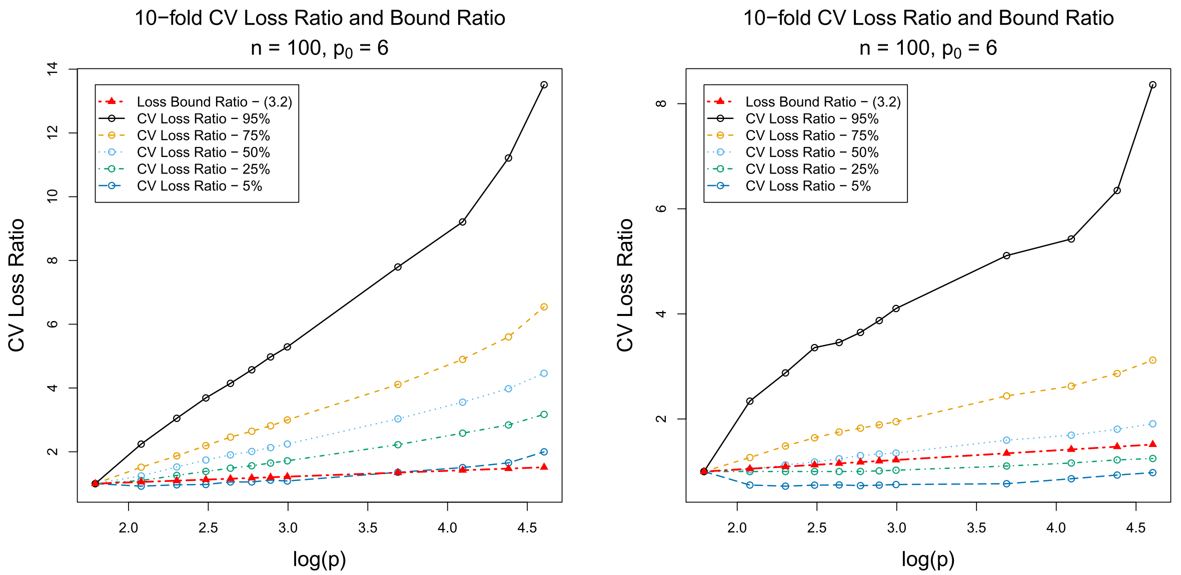
FIG. 4. *Average ratio of* λ^0 *to the observed selected* λ^*_p *over* 1000 *realizations as a function of* p *for* $n = 100$ *plotted on a log-scale. The "High SNR" and "Low SNR" settings correspond to* $\sigma^2 = 4$ *and* $\sigma^2 = 400$, *respectively.*

than the bounds imply, this further suggests that the deterioration gets worse for larger samples.

3.2 Independent Predictors

Here, we again assume that $\mathbf{y}$ is generated from the model given by (3.4) except in this section the columns of $\mathbf{X}$ 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.

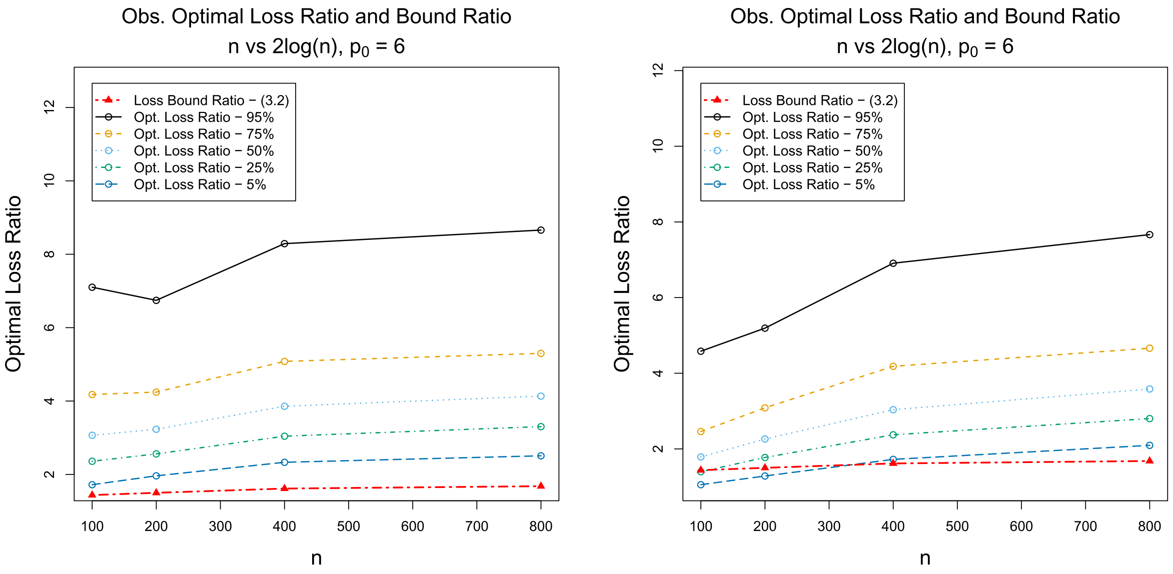
Figure 7 compares the percentiles of the optimal and CV loss ratios over 1000 realizations to the optimal



(a) High SNR

(b) Low SNR

FIG. 5. *Percentiles of the CV loss ratios 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.*



(a) High SNR (b) Low SNR

FIG. 6. *Percentiles of the optimal loss ratios 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.*

loss ratio suggested by the bound (3.2). We vary p from six to 1000, and denote the point where $p = n$ by the vertical line. In all four plots, the loss ratios predicted by the bound typically under-estimate the observed optimal and CV loss ratios. As in the orthogonal design case, these plots show that the bound does not adequately measure the deterioration in performance, and that the optimal and practical performance of the Lasso are sensitive to the number of predictor variables. These plots further indicate that deterioration occurs when $p > n$, though the deterioration pattern is less well-behaved.

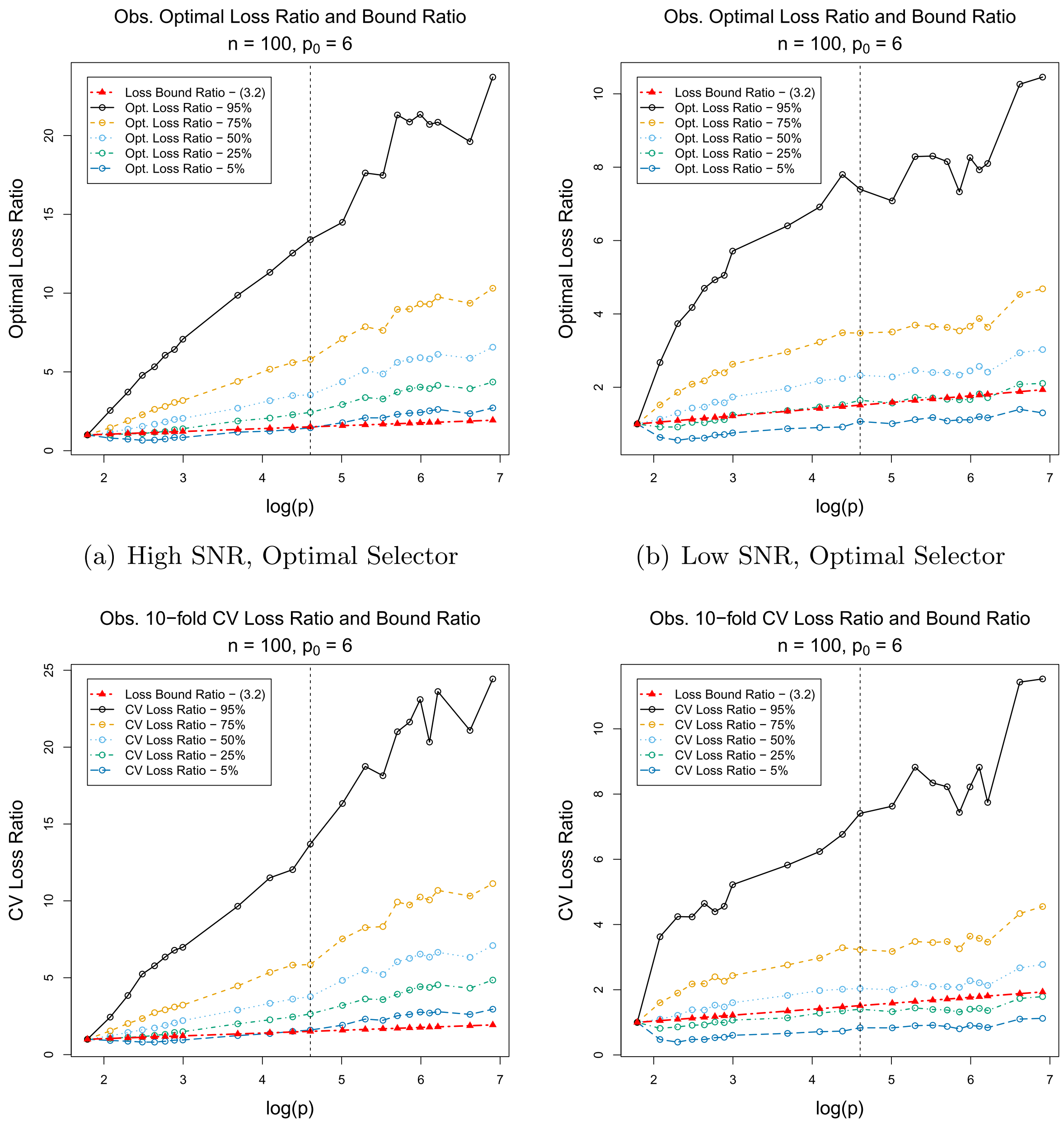
4. REAL DATA ANALYSIS

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

As an example of this, Bien, Taylor and Tibshirani (2013) 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 observations (n) and the number of mutation sites (p) for each dataset are listed in Table 2.

In their analysis, Bien, Taylor and Tibshirani (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 average predictive square error (APSE), which is defined as the average squared error between the values of the dependent variable on the test set and the values predicted by the model fit to the training set. We then study the APSE ratio, which compares the optimal APSE for APL to the optimal APSE for MEL. It is important to note that both the numerator and denominator in the APSE ratio include additional terms that depend on the noise term, which are not included



(a) High SNR, Optimal Selector

(b) Low SNR, Optimal Selector

(c) High SNR, 10-fold CV

(d) Low SNR, 10-fold CV

FIG. 7. *Percentiles of the optimal and CV loss ratios 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, *and the vertical line indicates the point where* $p = n$. *The "High SNR" and "Low SNR" settings correspond to* $\sigma^2 = 9$ *and* $\sigma^2 = 625$, *respectively.*

in the loss. Thus, the loss in estimation precision can be less apparent. To exemplify this, Appendix B studies the optimal APSE ratio in the context of the independent predictors example given in Section 3.2.

Figure 8 plots the ratios of the minimum test-set APSE obtained using the APL to that obtained using the MEL based on 20 random splits of the data for each of the six drugs.

TABLE 2
The number of observations 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

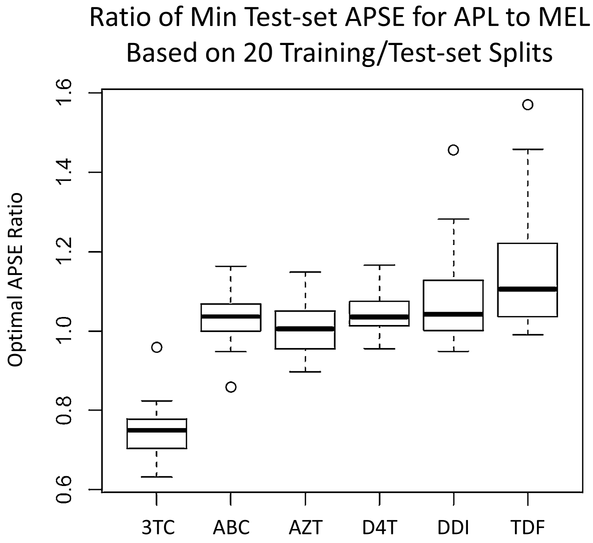


FIG. 8. *The ratio of the minimum test-set APSE obtained using the APL to that obtained using the MEL based on* 20 *random splits of the data for each of the six drugs.*

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.

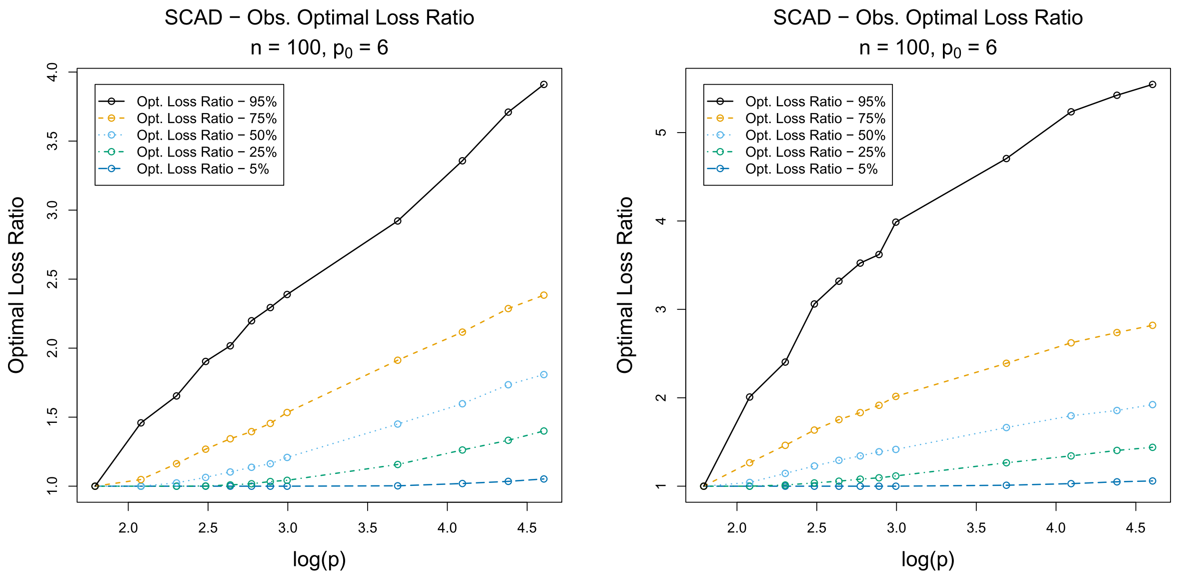
On the other hand, in four of the five remaining drugs—"ABC," "D4T," "DDI" and "TDF"—the inclusion of all pairwise interactions results in a significant deterioration in performance. Here, significance is determined using a Wilcoxon signed-rank test performed at a 0.05 significance level. Thus, although 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 loss bounds in the literature would suggest. We have proven 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 with one true predictor and an orthonormal deterministic design matrix. Our empirical results suggest that this deterioration persists as the sample size increases, and carries over to more general contexts.

In classical all-subsets regression, deterioration in the optimal loss does not occur, because it is always possible to recover the estimated true model while ignoring the extraneous predictors. This is not possible with the Lasso, because the only way to exclude extraneous predictors is to increase the amount of regularization imposed on all the estimated coefficients. This property is not unique to the Lasso, and preliminary results suggest that deterioration also occurs when using other regularization procedures. For example, Figure 9 plots the percentiles of the optimal loss ratios for SCAD (Fan and Li, 2001) under the set-up of Section 3.1 with orthogonal predictors. In both plots, there is evidence of deterioration. However, comparing Figure 9 to Figure 2, the degree of deterioration is typically less severe for SCAD than for the Lasso, especially in the High SNR setting. This partly due to the fact that the SCAD penalty imposes less shrinkage on the estimated coefficients. In the context of categorical predictors, Flynn, Hurvich and Simonoff (2016) also found evidence of deterioration when working with the group Lasso and the ordinal group Lasso. However, since the group Lasso and the ordinal group Lasso both impose more structure on the estimated coefficients, they reduce the effective degrees of freedom and the resulting observed deterioration for both methods is typically less severe than the deterioration observed when using the ordinary Lasso.

In light of the deterioration in performance, data analysts should be careful when using the Lasso and other regularization procedures as variable selection and estimation tools with high-dimensional data sets. One possible modification is to use the regularization procedure as a subset selector, but not as an estimation procedure. One 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 10 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 that imposes no shrinkage



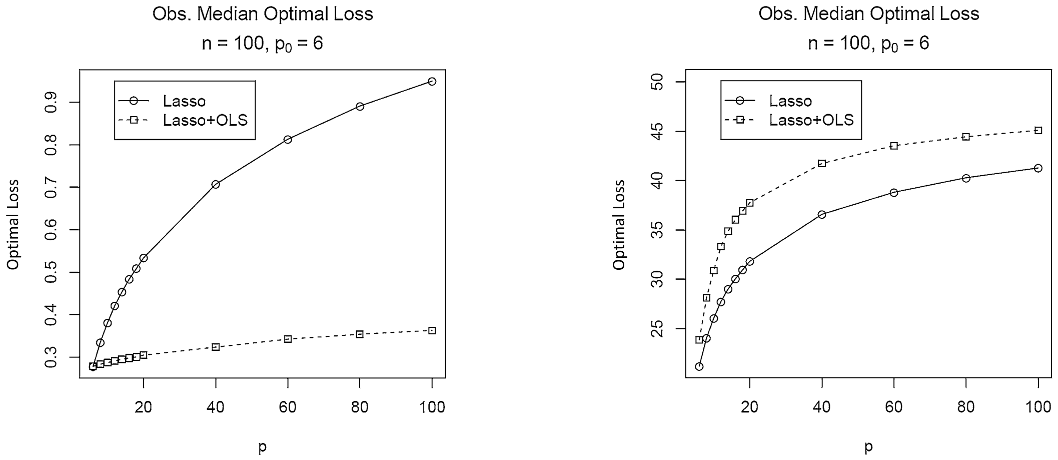
(a) High SNR (b) Low SNR

FIG. 9. *Percentiles of the optimal loss ratios for SCAD 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.*

on the estimated coefficients 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.

One may also consider alternatives to regularization. For example, Ando and Li (2014) achieved good performance in high-dimensional regression problems using a simple model averaging technique. More recently, Bertsimas, King and Mazumder (2016) developed a Mixed Integer Optimization approach to best subset selection, which they found could outperform the Lasso in numerical experiments. Further investigation into all of these techniques is an interesting area for future research.



(c) High SNR (d) Low SNR

FIG. 10. *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.*

APPENDIX A: TECHNICAL RESULTS

In this Appendix, we provide the proofs for the theoretical results presented in Section 2.

First, we prove the results for the more general p_0-sparse case.

PROOF OF LEMMA 2.1. First, note that $nL_p(\lambda_p^*) \leq \sum_{j=1}^{p_0} \beta_j^2$, because for any $\lambda_p^* \geq \max_{1\leq j\leq p} |z_j|$, $nL_p(\lambda_p^*) = \sum_{j=1}^{p_0} \beta_j^2$. If $\lambda_{p_0}^* \geq \max_{1\leq j\leq p_0} |z_j|$, then $nL_{p_0}(\lambda_{p_0}^*) = \sum_{j=1}^{p_0} \beta_j^2$ and the optimal λ will be one such that all of the estimated coefficients equal zero. No deterioration will occur in this case.

For the remainder of the proof, assume that $\lambda_{p_0}^* < \max_{1\leq j\leq p_0} |z_j|$. In this case, $nL_{p_0}(\lambda_{p_0}^*) < \sum_{j=1}^{p_0} \beta_j^2$. Consider

$$\begin{aligned} nL_p(\lambda_p^*) &= nL_{p_0}(\lambda_p^*) + \sum_{j=p_0+1}^{p} (|z_j| - \lambda_p^*)_+^2 \\ &\geq nL_{p_0}(\lambda_p^*) \geq nL_{p_0}(\lambda_{p_0}^*). \end{aligned}$$

The optimal loss does not deteriorate when equality holds. This occurs if and only if $\lambda_{p_0}^* = \lambda_p^*$ and $nL_p(\lambda_{p_0}^*) = nL_{p_0}(\lambda_{p_0}^*)$.

If $\lambda_{p_0}^* \geq \max_{p_0<j\leq p} |z_j|$, then

$$\begin{aligned} nL_{p_0}(\lambda_{p_0}^*) &= nL_{p_0}(\lambda_{p_0}^*) + \sum_{j=p_0+1}^{p} (|z_j| - \lambda_{p_0}^*)_+^2 \\ &= nL_p(\lambda_{p_0}^*). \end{aligned}$$

This implies that $nL_p(\lambda_p^*) = nL_{p_0}(\lambda_{p_0}^*)$ and no deterioration occurs.

Alternatively, if $\lambda_{p_0}^* < \max_{p_0<j\leq p} |z_j|$, then

$$\begin{aligned} nL_{p_0}(\lambda_{p_0}^*) &< nL_{p_0}(\lambda_{p_0}^*) + \sum_{j=p_0+1}^{p} (|z_j| - \lambda_{p_0}^*)_+^2 \\ &= nL_p(\lambda_{p_0}^*), \end{aligned}$$

so the optimal loss deteriorates.

It follows that deterioration occurs if and only if $\lambda_{p_0}^* < \max_{1\leq j\leq p_0} |z_j|$ and $\lambda_{p_0}^* < \max_{p_0<j\leq p} |z_j|$. □

PROOF OF THEOREM 2.1. By Lemma 2.1,

$$\begin{aligned} &\Pr\biggl(\frac{L_p(\lambda_p^*)}{L_{p_0}(\lambda_{p_0}^*)} > 1\biggr) \\ &\quad = \Pr\Bigl(\lambda_{p_0}^* \leq \max_{1\leq j\leq p_0} |z_j|, \lambda_{p_0}^* \leq \max_{p_0<j\leq p} |z_j|\Bigr) \\ &\quad \geq \Pr\Bigl(0 \leq \max_{1\leq j\leq p_0} |z_j|, 0 \leq \max_{p_0<j\leq p} |z_j|, \lambda_{p_0}^* = 0\Bigr) \\ &\quad = \Pr(\lambda_{p_0}^* = 0). \end{aligned}$$

Therefore, it is sufficient to show that $\Pr(\lambda_{p_0}^* = 0) > 0$ to show that the probability of deterioration is nonzero.

Consider the set

$$\begin{aligned} \mathcal{S} \equiv \{z_j, 1 \leq j \leq p_0 : \\ \beta_1 > z_1 > \beta_2 > z_2 > \cdots > \beta_{p_0} > z_{p_0} > 0\}. \end{aligned}$$

Assume that $z_j, 1 \leq j \leq p_0 \in \mathcal{S}$. This implies that $\sum_{j=1}^{k} (\beta_j - z_j) > 0$ for all $1 \leq k \leq p_0$.

For any $\lambda \in [0, z_{p_0})$,

$$nL_{p_0}(\lambda) = \sum_{j=1}^{p_0} (\beta_j - (z_j - \lambda))^2,$$

and

$$\frac{\partial nL_{p_0}(\lambda)}{\partial \lambda} = 2\sum_{j=1}^{p_0} (\beta_j - z_j) + 2p_0\lambda.$$

Since the derivative is an increasing function of λ and it is nonnegative at $\lambda = 0$, the minimum occurs at $\lambda = 0$.

Next, for any $1 < k \leq p_0$, consider $\lambda \in I_k = (z_k, z_{k-1})$. Over this interval,

$$nL_{p_0}(\lambda) = \sum_{j=1}^{k-1} (\beta_j - (z_j - \lambda))^2 + \sum_{j=k}^{p_0} \beta_j^2,$$

and

$$\frac{\partial nL_{p_0}(\lambda)}{\partial \lambda} = 2\sum_{j=1}^{k-1} (\beta_j - z_j) + 2(k-1)\lambda.$$

Since the derivative is an increasing function of λ and it is nonnegative at $\lambda = z_k$, the minimum occurs at $\lambda = z_k$. However, for any $1 < k \leq p_0$,

$$\begin{aligned} nL_{p_0}(z_k) &= \sum_{j=1}^{k-1} (\beta_j - (z_j - z_k))^2 + \sum_{j=k}^{p_0} \beta_j^2 \\ &> \sum_{j=1}^{p_0} (\beta_j - z_j)^2 = nL_{p_0}(0). \end{aligned}$$

Thus, $\lambda_{p_0}^* \notin I_k$ for any $1 < k \leq p_0$.

Finally, for any $\lambda \in [z_1, \infty)$,

$$nL_{p_0}(\lambda) = \sum_{j=1}^{p_0} \beta_j^2 > \sum_{j=1}^{p_0} (\beta_j - z_j)^2 = nL_{p_0}(0).$$

It follows that $\lambda_{p_0}^* = 0$ on $\mathcal{S}$.

Since the z_j's, $1 \leq j \leq p_0$ are independent normal random variables, it follows that

$$\Pr(\lambda_{p_0}^* = 0) \geq \Pr(\mathcal{S}) > 0.$$

Thus, equation (2.4) is satisfied. □

Next, to prove Theorem 2.2, we establish the following four lemmas. 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 A.1. *If* $\operatorname{sgn}(\beta_1) \neq \operatorname{sgn}(z_1)$, *then* $nL_p(\lambda_p^*) = \beta_1^2$ *for all* $0 < p \leq n$.

PROOF. If $\operatorname{sgn}(\beta_1) \neq \operatorname{sgn}(z_1)$, then for any $\lambda < \max_{1\leq j\leq p}|z_j|$,

$$\begin{aligned} nL_p(\lambda) &= \big(\beta_1 - \operatorname{sgn}(z_1)(|z_1| - \lambda)_+\big)^2 \\ &\quad + \sum_{j=2}^{p}(|z_j| - \lambda)_+^2 \\ &\geq \beta_1^2 + \sum_{j=2}^{p}(|z_j| - \lambda)_+^2 \geq \beta_1^2. \end{aligned}$$

Thus, $nL_p(\lambda_p^*) = \beta_1^2$. □

Lemma A.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 A.2. *If* $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(z_1)$, *then*

$$nL_1(\lambda_1^*) = \begin{cases} 0, & \textit{if } |\beta_1| \leq |z_1|, \\ (\beta_1 - z_1)^2, & \textit{otherwise.} \end{cases}$$

PROOF. Without loss of generality, assume that $\beta_1 > 0$ and, therefore, $z_1 > 0$. Consider

$$nL_1(\lambda) = \big(\beta_1 - (z_1 - \lambda)_+\big)^2.$$

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 $\lambda \in I$,

$$\frac{\partial nL_1(\lambda)}{\partial\lambda} = 2\big(\beta_1 - (z_1 - \lambda)\big).$$

Since the derivative is an increasing function of λ, a minimum occurs at $\lambda = 0$ if the derivative is nonnegative 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,

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

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

$$nL_1(\lambda_1^*) = \begin{cases} 0, & \text{if } 0 \leq \beta_1 < z_1, \\ (\beta_1 - z_1)^2, & \text{if } z_1 \leq \beta_1. \end{cases} \qquad \square$$

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 A.3. *Assume that* $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(z_1)$. *If* $\max_{2\leq j\leq p}|z_j| < |z_1|$, *then* $L_p(\lambda_p^*) = L_1(\lambda_1^*)$ *if and only if* $|\beta_1| < |z_1| - \max_{2\leq j\leq p}|z_j|$.

PROOF. Without loss of generality, assume that $\beta_1 > 0$ and, therefore, $z_1 > 0$. Also assume that $|z_2| > \cdots > |z_p|$. Consider

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

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 λ,

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

Since the derivative is a strictly increasing function of λ, a minimum occurs at $\lambda = 0$ if the derivative is nonnegative 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. Define

$$\lambda_I^* \equiv \operatorname*{argmin}_{\lambda\in I} nL_p(\lambda).$$

It follows that

$$\lambda_I^* = \begin{cases} z_1 - \beta_1, \quad \text{if } 0 < \beta_1 \le z_1 - |z_2|, \\ \dfrac{\sum_{j=1}^k |z_j| - \beta_1}{k}, \\ \quad \text{if } \sum_{j=1}^k |z_j| - k|z_k| < \beta_1 \le \sum_{j=1}^k |z_j| - k|z_{k+1}|, \\ \quad \text{for } k = 2, \dots, p-1, \\ \dfrac{\sum_{j=1}^p |z_j| - \beta_1}{p}, \\ \qquad \text{if } \sum_{j=1}^p |z_j| - p|z_p| < \beta_1 \le \sum_{j=1}^p |z_j|, \\ 0, \qquad \text{if } \sum_{j=1}^p |z_j| < \beta_1. \end{cases}$$

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

$$nL_p(\lambda_p^*) = \min(\beta_1^2, nL_p(\lambda_I^*)).$$

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

$$nL_p(\lambda_I^*) > (|z_2| - \lambda_I^*)^2 > 0 = nL_1(\lambda_1^*).$$

Similarly, if $z_1 < \beta_1$, then either $\lambda_I^* > 0$ so that

$$nL_p(\lambda_I^*) > (\beta_1 - (z_1 - \lambda_I^*))^2 > (\beta_1 - z_1)^2 = nL_1(\lambda_1^*),$$

or $\lambda_I^* = 0$ and

$$\begin{aligned} nL_p(\lambda_I^*) &= (\beta_1 - z_1)^2 + \sum_{j=2}^p |z_j|^2 > (\beta_1 - z_1)^2 \\ &= nL_1(\lambda_1^*). \end{aligned}$$

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

LEMMA A.4. *Assume that* $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(z_1)$. *If* $\max_{2 \le j \le p} |z_j| > |z_1|$, *then* $L_p(\lambda_p^*) > L_1(\lambda_1^*)$ *for all* $\beta_1 \ne 0$.

PROOF. Without loss of generality, assume that $\beta_1 > 0$ and, therefore, $z_1 > 0$. Also assume that $|z_2| > \cdots > |z_p|$. Consider

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

Define

$$\tilde{k} = \max_{2 < k \le p} \{k : |z_k| > z_1\}.$$

The derivative of $nL_p(\lambda)$ does not exist at $\lambda = z_1$. However, by a similar argument to that used in the proof of Lemma A.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 λ, we consider the intervals $I_1 = [0, z_1)$, $I_2 = (z_1, |z_2|]$, and $I_3 = (|z_2|, \infty)$ separately. Define

$$\lambda_{I_j}^* = \operatorname*{argmin}_{\lambda \in I_j} nL_p(\lambda)$$

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 λ, a minimum occurs at $\lambda = 0$ if the derivative is nonnegative 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 $0 < \beta_1 \le \sum_{j=1}^{\tilde{k}} |z_j| - \tilde{k} z_1$. Otherwise, a minimum occurs at a point where the derivative is zero. It follows that

$$\lambda_{I_1}^* = \begin{cases} z_1, \quad \text{if } 0 < \beta_1 \le \sum_{j=1}^{\tilde{k}} |z_j| - \tilde{k} z_1, \\ \dfrac{\sum_{j=1}^{\tilde{k}} |z_j| - \beta_1}{\tilde{k}}, \\ \quad \text{if } \sum_{j=1}^{\tilde{k}} |z_j| - \tilde{k} z_1 < \beta_1 \le \sum_{j=1}^{\tilde{k}} |z_j| - \tilde{k}|z_{\tilde{k}+1}|, \\ \dfrac{\sum_{j=1}^k |z_j| - \beta_1}{k}, \\ \quad \text{if } \sum_{j=1}^k |z_j| - k|z_k| < \beta_1 \le \sum_{j=1}^k |z_j| - k|z_{k+1}|, \\ \quad \text{for } k = \tilde{k} + 1, \dots, p-1, \\ \dfrac{\sum_{j=1}^p |z_j| - \beta_1}{p}, \\ \quad \text{if } \sum_{j=1}^p |z_j| - p|z_p| < \beta_1 \le \sum_{j=1}^p |z_j|, \\ 0, \quad \text{if } \sum_{j=1}^p |z_j| < \beta_1. \end{cases}$$

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 $nL_p(\lambda^*_{I_2}) = \beta_1^2$.

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

$$nL_p(\lambda^*_p) = \min(\beta_1^2, nL_p(\lambda^*_{I_1})).$$

By a similar argument to that used in the proof of Lemma A.3, it follows that $nL_p(\lambda^*_p) > nL_1(\lambda^*_1)$. □

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 A.3 and A.4, when the sign of z_1 is correct, $L_p(\lambda^*_p) > L_1(\lambda^*_1)$ unless $|\beta_1| < |z_1| - \max_{2\le j\le p}|z_j|$.

PROOF OF THEOREM 2.2. By Lemma A.1,

$$\begin{aligned}
&\Pr(nL_p(\lambda^*_p) > nL_1(\lambda^*_1))\\
&\quad = \Pr(nL_p(\lambda^*_p) > nL_1(\lambda^*_1) | \operatorname{sgn}(z_1) = \operatorname{sgn}(\beta_1))\\
&\qquad \cdot \Pr(\operatorname{sgn}(z_1) = \operatorname{sgn}(\beta_1)).
\end{aligned}$$

Without loss of generality, assume that $\beta_1 > 0$. By Lemmas A.3–A.4, this is equal to

$$\begin{aligned}
&(1 - \Pr(nL_p(\lambda^*_p)\\
&\quad = nL_1(\lambda^*_1) | z_1 > 0))\Pr(z_1 > 0)\\
&\quad = \Bigl(1 - \Pr\Bigl(\beta_1 < z_1 - \max_{2\le j\le p}|z_j| \,|\, z_1 > 0\Bigr)\Bigr)\\
&\qquad \cdot \Pr(z_1 > 0).
\end{aligned}$$

We can evaluate these probabilities explicitly. First, consider

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

Next,

$$\begin{aligned}
&\Pr\Bigl(\beta_1 < z_1 - \max_{2\le j\le p}|z_j| \,|\, z_1 > 0\Bigr)\\
&\quad = \frac{\Pr(\{\beta_1 < z_1 - \max_{2\le j\le p}|z_j|\} \cap \{z_1 > 0\})}{\Pr(z_1 > 0)}\\
&\quad = \frac{\Pr(\beta_1 < z_1 - \max_{2\le j\le p}|z_j|)}{\Pr(z_1 > 0)},
\end{aligned}$$

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

$$\begin{aligned}
&\Pr\Bigl(\beta_1 < z_1 - \max_{2\le j\le p}|z_j|\Bigr)\\
&\quad = \Pr\Bigl(\bigcap_{j=2}^{p}\{\beta_1 < z_1 - |z_j|\}\Bigr)\\
&\quad = \int_{z_1=\beta_1}^{z_1=\infty}\Bigl[\int_{z_2=-(z_1-\beta_1)}^{z_2=z_1-\beta_1} f_2(z_2)\,\mathrm{d}z_2\Bigr]^{p-1} f_1(z_1)\,\mathrm{d}z_1\\
&\quad = \int_{\beta_1}^{\infty}\Bigl[2\Phi\Bigl(\frac{z_1-\beta_1}{\sigma}\Bigr) - 1\Bigr]^{p-1} f_1(z_1)\,\mathrm{d}z_1\\
&\quad = \frac{1}{\sigma}\int_{\beta_1}^{\infty}\Bigl[2\Phi\Bigl(\frac{z_1-\beta_1}{\sigma}\Bigr) - 1\Bigr]^{p-1} \phi\Bigl(\frac{z_1-\beta_1}{\sigma}\Bigr)\,\mathrm{d}z_1,
\end{aligned}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the probability distribution functions (p.d.f.) of z_1 and z_2, respectively, and $\phi(\cdot)$ is the p.d.f. of the standard normal distribution. Substituting

$$w = 2\Phi\Bigl(\frac{z_1-\beta_1}{\sigma}\Bigr) - 1,$$

$$\Pr\Bigl(\beta_1 < z_1 - \max_{2\le j\le p}|z_j|\Bigr) = \frac{1}{2}\int_0^1 w^{p-1}\,\mathrm{d}w = \frac{1}{2p}.$$

Thus,

$$\Pr\Bigl(\beta_1 < z_1 - \max_{2\le j\le p}|z_j| \,|\, z_1 > 0\Bigr) = \frac{\frac{1}{2p}}{\Phi(\frac{\beta_1}{\sigma})}. \tag{A.2}$$

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

$$\Pr(L_p(\lambda^*_p) > L_1(\lambda^*_1)) = \Phi\Bigl(\frac{\beta_1}{\sigma}\Bigr) - \frac{1}{2p}.$$ □

Lastly, we provide the proof for Theorem 2.3.

PROOF OF THEOREM 2.3. Without loss of generality, assume that $\beta_1 > 0$. Define

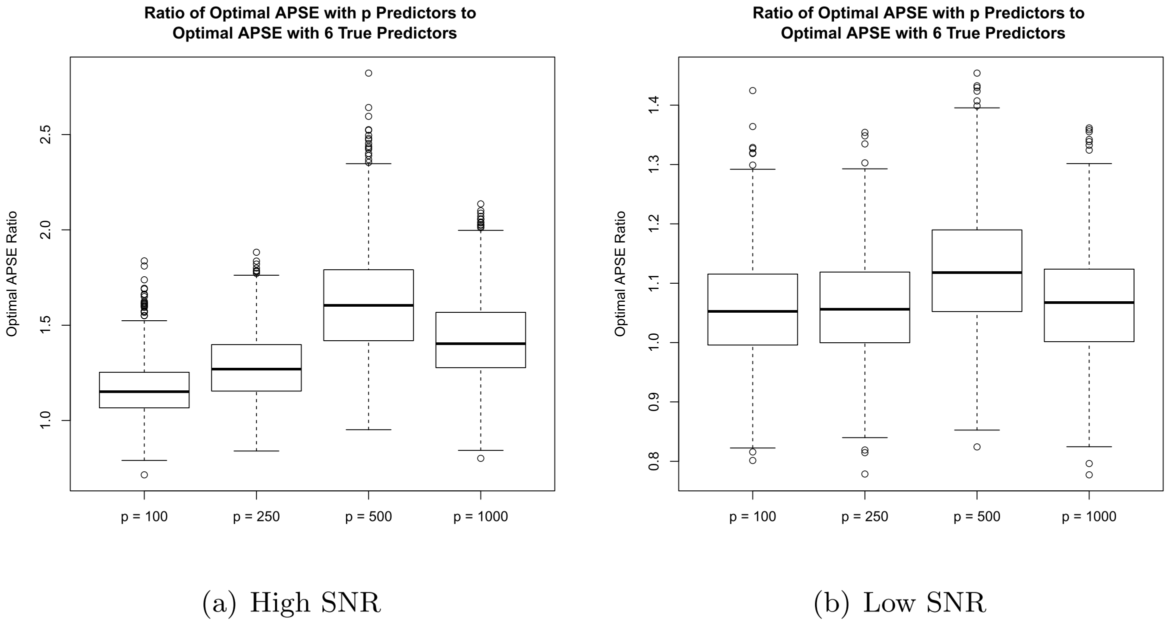
$$\mathcal{A} := \{z_1, z_2 : z_2 > z_1 > \beta_1\}.$$

Note that $\Pr((z_1, z_2) \in \mathcal{A}) > 0$. By Lemmas A.2 and A.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

$$\begin{aligned}
&\mathrm{E}\Bigl(\frac{L_p(\lambda^*_p)}{L_1(\lambda^*_1)}\Bigr)\\
&\quad = \int_{z_p=-\infty}^{z_p=\infty}\cdots\int_{z_1=-\infty}^{z_1=\infty}\frac{L_p(\lambda^*_p)}{L_1(\lambda^*_1)}\,\mathrm{d}f_1(z_1)\cdots\mathrm{d}f_p(z_p)
\end{aligned}$$



(a) High SNR

(b) Low SNR

FIG. A.1. *Optimal APSE 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.*

$$\geq \int_{z_p=-\infty}^{z_p=\infty} \cdots \iint_{(z_1,z_2)\in\mathcal{A}} \frac{L_p(\lambda_p^*)}{L_1(\lambda_1^*)}\,\mathrm{d}f_1(z_1)\cdots\mathrm{d}f_p(z_p)$$

$$= \infty. \qquad \square$$

APPENDIX B: OPTIMAL APSE RATIO

Here, we return to the independent predictors example in Section 3.2. To study the behavior of the optimal APSE, we evaluate the APSE for each realization on a simulated test set. Figure A.1 presents boxplots of the ratios of the estimated optimal APSE with p predictors to the estimated optimal APSE with the six true predictors where p is taken to be 100, 250, 500 and 1000 and $n = 100$. A comparison of this figure to the median optimal loss ratios presented in Figure 7 demonstrates that while deterioration is still observed, the optimal APSE ratios can be smaller than the optimal loss ratios. To understand why this is the case, note that the APSE is equal to

$$\frac{1}{n}\|\mathbf{y}^* - \hat{\mathbf{y}}\|^2 = \frac{1}{n}(\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\|^2 + 2(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})^T\varepsilon^* + \|\varepsilon^*\|^2),$$

where $\cdot^*$ is with respect to an independent test set. Thus, the optimal APSE ratios can be smaller than the optimal loss ratios due to the presence of additional terms in both the numerator and denominator of the APSE ratio.

These figures also suggest that the deterioration pattern is less well-behaved when $p > n$ than it is when $p < n$, which is consistent with the results found in Section 3.2.

ACKNOWLEDGMENTS

Cheryl J. Flynn's work was done while the author was at New York University.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR* 267–281.

ANDO, T. and LI, K.-C. (2014). A model-averaging approach for high-dimensional regression. *J. Amer. Statist. Assoc.* **109** 254–265. MR3180561

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A Lasso for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*. Springer, Berlin. MR2807761

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via l_1 penalized least squares. In *Learning Theory. Lecture Notes in Computer Science* **4005** 379–391. Springer, Berlin. MR2280619

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007a). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007b). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194.

CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982

CRAVEN, P. and WAHBA, G. (1978). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FLYNN, C. J., HURVICH, C. M. and SIMONOFF, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *J. Amer. Statist. Assoc.* **108** 1031–1043. MR3174682

FLYNN, C. J., HURVICH, C. M. and SIMONOFF, J. S. (2016). Deterioration of performance of the Lasso with many predictors: Discussion of a paper by Tutz and Gertheiss. *Stat. Model.* **16** 212–216.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GREENSHTEIN, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.* **34** 2367–2386. MR2291503

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

HOMRIGHAUSEN, D. and MCDONALD, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Mach. Learn.* **97** 65–78. MR3252827

HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020

HYNDMAN, R. J., BOOTH, H. and YASMEEN, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography* **50** 261–283.

LENG, C., LIN, Y. and WAHBA, G. (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica* **16** 1273–1284. MR2327490

LIN, D., FOSTER, D. P. and UNGAR, L. H. (2011). VIF regression: A fast regression algorithm for large data. *J. Amer. Statist. Assoc.* **106** 232–247. MR2816717

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990

RHEE, S.-Y., TAYLOR, J., WADHERA, G., BEN-HUR, A. and BRUTLAG, D. L. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA* **103** 17355–17360.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

THRAMPOULIDIS, C., PANAHI, A. and HASSIBI, B. (2015). Asymptotically exact error analysis for the generalized l_2^2-LASSO. Preprint. Available at arXiv:1502.06287.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VIDAURRE, D., BIELZA, C. and LARRAÑAGA, P. (2013). A survey of L_1 regression. *Int. Stat. Rev.* **81** 361–387. MR3146024

YU, Y. and FENG, Y. (2014). Modified cross-validation for penalized high-dimensional linear regression models. *J. Comput. Graph. Statist.* **23** 1009–1027. MR3270708

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967