The lasso, persistence, and cross-validation

Darren Homrighausen DARRENHO@STAT.COLOSTATE.EDU
Department of Statistics, Colorado State University, Fort Collins, CO 80523

Daniel J. McDonald DAJMCDON@INDIANA.EDU
Department of Statistics, Indiana University, Bloomington, IN 47408

Abstract

During the last fifteen years, the lasso procedure has been the target of a substantial amount of theoretical and applied research. Correspondingly, many results are known about its behavior for a fixed or optimally chosen smoothing parameter (given up to unknown constants). Much less, however, is known about the lasso's behavior when the smoothing parameter is chosen in a data dependent way. To this end, we give the first result about the risk consistency of lasso when the smoothing parameter is chosen via cross-validation. We consider the high-dimensional setting wherein the number of predictors $p = n^\alpha$, $\alpha > 0$ grows with the number of observations.

1. Introduction

Since its introduction in the statistical (Tibshirani, 1996) and signal processing (Chen et al., 1998) communities, ℓ_1-penalized linear regression has been a fixture as both a data analysis tool and as a subject for deep theoretical investigations. In particular, for a response vector $Y \in \mathbb{R}^n$, design matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, and tuning parameter t, we consider the lasso problem of finding

$$\widehat{\beta}_t \in \underset{\beta \in \mathcal{B}_t}{\operatorname{argmin}} \frac{1}{n} ||Y - \mathbb{X}\beta||_2^2 \tag{1}$$

where $\mathcal{B}_t := \{\beta : ||\beta||_1 \leq t\}$ and $||\cdot||_2$ and $||\cdot||_1$ indicate the Euclidean and ℓ_1-norms respectively. By convexity, for each t, there is always at least one solution in equation (1). Note that, while the solution is not necessarily unique (i.e.: when $\operatorname{rank}(\mathbb{X}) < p$), this detail is unimportant for our purposes and we abuse notation slightly by referring to $\widehat{\beta}_t$ as 'the' lasso solution.

There is a large and growing literature investigating the asymptotic properties of the lasso solution. We highlight some results here, but it is not our intention to give an exhaustive overview of the field. In one of the earliest theoretical papers, Fu & Knight (2000) examine the asymptotic distribution of the lasso solution under the assumption that the sample covariance matrix has a nonnegative definite limit and p is fixed. Alternatively, Zou (2006); Wainwright (2009); Donoho et al. (2006); Meinshausen & Yu (2009); Meinshausen & Bühlmann (2006) and Zhao & Yu (2006) have investigated the model selection properties of the lasso. These results, which hold under various sparsity and "irrepresentability" conditions, show that if we assume the best predicting linear model to be sparse, the lasso will tend to asymptotically recover those predictors.

The criterion we focus on for this paper is risk consistency, alternatively known as persistence. That is, we require the prediction risk of the estimated model to converge to that of the best linear oracle predictor. Risk consistency has previously been investigated by Bunea et al. (2007), van de Geer (2008), and Greenshtein & Ritov (2004). These results depend critically on the choice of tuning parameters and are typically of the form: if $t = t_n$ is such that $t_n = o(a_n)$ for some rate a_n, such as $a_n = (n/\log(n))^{1/4}$, then $\widehat{\beta}_{t_n}$ is persistent. However comforting results of this type are, this theoretical guidance says little about the properties of the lasso when the tuning parameter is chosen in a data-dependent, and hence stochastic, way.

There are several proposed techniques for choosing t, or equivalently, the parameter in the Lagrangian formulation, commonly denoted by λ. Zou et al. (2007) and Tibshirani & Taylor (2012) investigate using the "degrees of freedom" of a lasso solution which can be informally defined as the trace of the covariance between the lasso solution and the response Y. The logic

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

of this procedure is that an unbiased estimator of the degrees of freedom provides an unbiased estimator of the risk. Hence, minimizing this estimator of the risk provides a method for choosing the tuning parameter. Another risk estimator is the adapted Bayesian information criterion of Wang & Leng (2007). It uses a plug-in estimator of the second-order Taylor's expansion of the risk.

However, in many papers, for example (Tibshirani, 1996; Greenshtein & Ritov, 2004; Hastie et al., 2009; Efron et al., 2004; Zou et al., 2007; Tibshirani, 2011; van de Geer & Lederer, 2011) and in the `R` package `glmnet` described by Friedman et al. (2010), the recommended or default technique for selecting t in the lasso problem is to choose $t = \widehat{t}$ such that $\widehat{t}$ minimizes the cross-validation (which we abbreviate CV) estimator of the risk.

The main contribution of this paper is to show that the use of cross-validation to choose the tuning parameter in lasso remains persistent relative to the theoretically optimal, but empirically unavailable, non-stochastic choice. We consider the high-dimensional regime where $p_n = n^\alpha$, for a positive α that is to be discussed in Section 3.

Some results supporting the use of CV for statistical algorithms other than lasso are known. For instance kernel regression (Györfi et al., 2002, Theorem 8.1), k-nearest neighbors (Györfi et al., 2002, Theorem 8.2), and neural networks (Plutowski et al., 1994) all behave well with tuning parameters selected via CV. However, the vast literature on the lasso is strangely silent on the theoretical behavior of the cross-validated estimator. The prevailing heuristic understanding of the performance of $\widehat{\beta}_{\widehat{t}}$, which is the lasso solution with t chosen by CV, is encapsulated in the statement

> Regarding the choice of the regularization parameter, we typically use $[\widehat{t}]$ from cross-validation. 'Luckily', empirical and some theoretical indications support [good performance]...(Tibshirani, 2011, Bühlmann's comments).

The supporting theory for non-lasso methods suggests that there should be corresponding theory for the lasso. However, other results are not so encouraging. In particular, Shao (1993) shows that cross-validation is inconsistent for model selection. As lasso implicitly does model selection, and shares many connections with forward stepwise regression (Efron et al., 2004), this raises a concerning possibility that lasso might similiarly be inconsistent for prediction under cross-validation. Likewise, Leng et al. (2006) show that using prediction accuracy (which is what cross-validation estimates) as a criterion for choosing the tuning parameter in lasso fails to recover the sparsity pattern consistently in an orthogonal design setting. Furthermore, Xu et al. (2008) show that sparsity inducing algorithms like lasso are not (uniformly) algorithmically stable. In other words, leave-one-out versions of the lasso estimator are not uniformly close to each other. As shown in Bousquet & Elisseeff (2002), algorithmic stability is a sufficient, but not necessary, condition for persistence.

These results taken as a whole leave the lasso in an unsatisfactory position, with some theoretical results and generally accepted practices advocating the use of cross-validation while others indicate that cross-validation may not be a sound method for selecting the tuning parameter at all.

In this paper, we show that the lasso under random design with cross-validated tuning parameter is indeed risk consistent under some conditions on the joint distribution of the design that generates $\mathbb{X}$ and the response Y. In Section 2, we outline the mathematical setup for the lasso prediction problem and discuss some empirical concerns. Section 3 contains the main result and associated conditions. Section 4 presents some useful lemmas and provides the proof of our results, while Section 5 summarizes our contribution.

2. Notation and definitions

2.1. Preliminaries

Suppose we observe pairs $Z_{i,n}^\top = (Y_{i,n}, X_{i,n}^\top)$ of predictor variables, $X_{i,n} \in \mathbb{R}^{p_n}$, and response variables, Y_i, where $Z_{i,n} \overset{i.i.d}{\sim} F_n$ for $i = 1, 2, \ldots, n$ and the distribution F_n is in some class $\mathcal{F}$ to be specified later. Here, we use the notation p_n to allow the number of predictor variables to change with n. For simplicity of notation, in what follows, we omit the subscript n when there is little risk of confusion.

We consider the problem of estimating the best linear functional $f(\mathcal{X}_1, \ldots, \mathcal{X}_p) = \beta^\top \mathcal{X}$ for predicting $\mathcal{Y}$, when $\mathcal{Z}^\top = (\mathcal{Y}, \mathcal{X}^\top) \sim F_n$ is a new data point from the same distribution and β is constrained to be in some set $\mathcal{B}$. We use the L^2-risk of a predictor $\beta = (\beta_1, \ldots, \beta_p)^\top$, defined as

$$L(\beta) := \mathbb{E}_{F_n}\left[(\mathcal{Y} - \beta^\top \mathcal{X})^2\right], \tag{2}$$

for our criterion. Note that the expectation here is taken only over the new datum $\mathcal{Z}$ and not over any observables which may or may not be used to choose β. This will be our convention throughout: $L(\cdot)$ denotes

the expectation over a new data point conditional on the original sample.

Using the n independent observations $Z_1, \ldots, Z_n$, we can form the response vector $Y := (Y_i)_{i=1}^n$ and design matrix $\mathbb{X} := [X_1, \ldots, X_n]^\top$. Then, given a vector β, we can write the squared-error objective function as

$$\widehat{L}(\beta) := \frac{1}{n}||Y - \mathbb{X}\beta||_2^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2. \quad (3)$$

Recall that we define $\mathcal{B}_t := \{\beta : ||\beta||_1 \leq t\}$. Analogously to equation (3), we write the *K-fold cross-validation estimator of the risk with respect to some regularization set* $\mathcal{B}_t$, which we abbreviate to CV-risk, as

$$\begin{aligned}\widehat{L}_{V_n}(t) &= \widehat{L}_{V_n}\left(\widehat{\beta}_t^{(v_1)}, \ldots, \widehat{\beta}_t^{(v_{K_n})}\right) \\ &:= \frac{1}{K}\sum_{v \in V_n}\frac{1}{|v|}\sum_{r \in v}\left(Y_r - X_r^\top \widehat{\beta}_t^{(v)}\right)^2. \quad (4)\end{aligned}$$

Here, $V_n = \{v_1, \ldots, v_K\}$ is a set of validation sets, $\widehat{\beta}_t^{(v)}$ is the estimator in equation (1) with the observations in the validation set v removed, and $|v|$ indicates the cardinality of the set v. Notice in particular that the CV-risk is a function of $\mathcal{B}_t$, and hence t, rather than a single predictor $\widehat{\beta}$. As the more complete notation makes clear, the CV-risk is actually a function of K different predictors $\widehat{\beta}_t^{(v)}$. Lastly, we define the CV-risk minimizing choice of tuning parameter to be

$$\widehat{t} = \operatorname*{argmin}_{t \in T_n} \widehat{L}_{V_n}(t). \quad (5)$$

2.2. Choosing the set T_n

In practice, an upper bound must be selected for any grid-search optimization over t. Note that more advanced optimization techniques are generally not practical as the CV objective function in equation (5) is often noisy. To define such an upper bound in a practical way, it should be large enough to include all possible estimators in a given class while still being finite. This implies we must choose T_n to be a function of the data, in the same way that $\widehat{t}$ is. The specifics of T_n depend on the regularizing set $\mathcal{B}_t$. This upper bound has a nontrivial impact on the quality of the recovery, as choosing a value too small may possibly eliminate the best solutions. Thus, treating the upper bound as a random function of the data is more realistic from a statistical practice point of view.

We note that, by the definition of $\mathcal{B}_t$, $\widehat{\beta}_t$ must be in the ℓ_1-ball with radius t. This constraint is only binding (Osborne et al., 2000) if

$$t < \min_{\eta \in \mathcal{K}} ||\widehat{\beta}^0 + \eta||_1 =: t_0,$$

where $\widehat{\beta}^0 := (\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top Y$ is a least squares solution, $(\cdot)^\dagger$ is a pseudoinverse, and $\mathcal{K} := \{a : \mathbb{X}a = 0\}$ is the null space of $\mathbb{X}$. Observe that $\mathcal{K} = \{0\}$ if $n \geq p$ and otherwise $\mathcal{K}$ has dimension $p - n$, and $\widehat{\beta}^0$ is not unique (both of these statements assume the columns of $\mathbb{X}$ are linearly independent). In either case, if $t \geq t_0$, then $\widehat{\beta}_t$ is equal to a least squares solution. Therefore, we define $T_n := [0, t_{\max}]$, where

$$t_{\max} := \left|\left|\widehat{\beta}^0\right|\right|_1,$$

and $\widehat{\beta}^0$ is the least squares solution when $(\cdot)^\dagger$ is given by the Moore-Penrose inverse rather than any other pseudoinverse. It holds that $t_{\max} \geq t_0$ as $0 \in \mathcal{K}$ and therefore $t_{\max}$ is a bit conservative in the sense that the lasso solution is already a least squares solution for values of $t \in [t_0, t_{\max})$. We make this definition to eliminate the explicit dependence on the null space of $\mathbb{X}$ that appears in the definition of t_0. As an aside, in the `R` implementation of LARS (Efron et al., 2004), a fraction of the ℓ_1-norm of the 'saturated' model is used, with a default fraction value of 1. This coincides with our choice of $t_{\max}$, the ℓ_1-norm of a particular least squares solution.

3. Main results

We define the oracle estimator generated by $\mathcal{B}_t$ to be

$$\beta_t := \operatorname*{argmin}_{\beta \in \mathcal{B}_t} L(\beta).$$

A natural criterion for studying the performance of the estimator $\widehat{\beta}_{\widehat{t}}$ is the *excess risk of* $\widehat{\beta}_{\widehat{t}}$ *relative to* β_t, which we define as

$$\mathcal{E}(\widehat{t}, t) := L\left(\widehat{\beta}_{\widehat{t}}\right) - L(\beta_t). \quad (6)$$

This criterion allows for meaningful theory when the oracle linear model is not risk consistent; that is, when the term $L(\beta_t)$ does not necessarily go to zero. This is particularly important in this case, as the conditional expectation of $\mathcal{Y}$ given $\mathcal{X}$ need not be even approximately linear. It is important to clarify two aspects of this definition of excess risk. First, $\mathcal{E}(\widehat{t}, t)$ is random due to the term $L\left(\widehat{\beta}_{\widehat{t}}\right)$ in equation (6). Here, $L\left(\widehat{\beta}_{\widehat{t}}\right)$ is a function of the data, and the expectation involved is only over a new test random variable $\mathcal{Z}$ and not with respect to the observed data used to choose either $\widehat{t}$ or $\widehat{\beta}_{\widehat{t}}$. Second, conventionally, $\mathcal{B}_{\widehat{t}} = \mathcal{B}_t$, and so the excess risk is necessarily nonnegative. However, as we are examining the case where the tuning parameter $\widehat{t}$ (and hence the optimization set $\mathcal{B}_{\widehat{t}}$) is estimated, $\mathcal{E}(\widehat{t}, t)$ may be negative. Note

that $\mathbb{P}(\mathcal{E}(\widehat{t},t)\leq 0)\leq \mathbb{P}(\mathcal{B}_t\subseteq\mathcal{B}_{\widehat{t}})=\mathbb{P}(t\leq\widehat{t})$, because β_t is the risk minimizer over all of $\mathcal{B}_t$. We return to this issue in the proof of our main result in the next section. We wish to show that the excess risk of an estimator with tuning parameter chosen by CV goes to zero in probability. First, we define the following set of distributions

Definition 3.1. *Let*

$$\mathcal{F}:=\Big\{(F_n)_{n\geq 1}\colon \exists C<\infty \text{ for all } n \text{ s.t. } \mathbb{E}_{F_n}\max_{0\leq j,k\leq p}(\mathcal{Z}_j\mathcal{Z}_k-\mathbb{E}_{F_n}\mathcal{Z}_j\mathcal{Z}_k)^2\leq C\Big\}.$$

Heuristically, $\mathcal{F}$ is the set of all triangular distributions where a universal constant exists that bounds the variance of each the $(p+1)^2$ interaction terms.

Remark 1. *Definition 3.1 is the same moment condition imposed in Greenshtein & Ritov (2004) to show risk consistency of the lasso in high-dimensional settings.*

We also state the following conditions.

Condition 1. *All $(F_n)\in\mathcal{F}$ are such that*

$$\mathbb{E}_{F_n}\left[t_{\max}^4\right]=o(t_n^4).$$

Condition 2. *For any cross-validation procedure V_n, which is defined in equation (4) , there exists a sequence of constants (c_n) such that for all $v\in V_n, |v|\geq c_n$. Additionally, for any $v\neq v'\in V_n$, $v\cap v'=\emptyset$.*

For example, with K-fold cross-validation, we can take $c_n=\lfloor n/K\rfloor$, which is the integer part of n/K.

Before explaining these conditions in depth, we state our main result.

Theorem 3.2. *Suppose $p_n=n^\alpha$ for some $\alpha>0$. Let $(F_n)\in\mathcal{F}$ be given and suppose Condition 1 and Condition 2 hold.*

Then, for any $\delta>0$,

$$\mathbb{P}(\mathcal{E}(\widehat{t},t_n)>\delta)=o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right).$$

This result shows that choosing the tuning parameter $\widehat{t}$ with CV and then estimating β by $\widehat{\beta}_{\widehat{t}}$ has the same asymptotic risk as minimizing the true risk over the set $\mathcal{B}_{t_n}$ as long as $c_n\asymp n$ which is the case for K-fold CV.

The inclusion of t_n in Theorem 3.2 deserves comment. Here, t_n is any sequence of non-random constants which determine the amount of regularization. As mentioned in Greenshtein & Ritov (2004), if t_n grows as fast or faster than $\left(\frac{n}{\log n}\right)^{1/4}$, then $L\left(\widehat{\beta}_{t_n}\right)-L\left(\beta_{t_n}\right)$ does not necessarily converge to 0 in probability and hence the lasso is not persistent. However, if $t_n=o((n/\log n)^{1/4})$, then the lasso is persistent. Therefore, we choose the oracle risk over $\mathcal{B}_{t_n}$ as our comparison for persistence.

We discuss Condition 1 next.

Explaining Condition 1 Some assumptions about the design distribution can be used to derive sufficient conditions for the moment condition we impose. Suppose that $p_n=n^\alpha$ for $\alpha>0$. Then

$$\begin{aligned}\mathbb{E}[t_{\max}^4]&=\mathbb{E}\left[\left|\left|\widehat{\beta}^0\right|\right|_1^4\right]\\&\leq\mathbb{E}\left[|||(X^\top X)^\dagger X^\top|||_1^4\,||Y||_1^4\right] \quad (7)\\&\leq p^2\mathbb{E}\left[|||(X^\top X)^\dagger X^\top|||_2^4\,||Y||_1^4\right]\\&=p^2\mathbb{E}_X\left[|||(X^\top X)^\dagger X^\top|||_2^4\right]\mathbb{E}_{Y|X}\left[||Y||_1^4\right]\\&=n^{2\alpha}\mathbb{E}_X\left[\sigma_{\min}^+(X)^{-4}\right]\mathbb{E}_{Y|X}\left[||Y||_1^4\right], \quad (8)\end{aligned}$$

where $\sigma_{\min}^+(A)$ is the smallest non-negative singular value of a matrix A, $|||\cdot|||_s$ is the operator norm corresponding to the ℓ_s vector norm, and equation (7) uses the sub-multiplicative property of the operator norm.

Suppose that our model is the usual nonparametric regression model $Y=m(X)+e$, where X and e are stochastically independent random variables and e has zero mean. If there exists a constant C independent of n such that $\operatorname{ess\,sup}_x m(x)<C$ with respect to the distribution of X and $\mathbb{E}e^4<\infty$, then there exists a $C'<\infty$, again independent of n, such that

$$\begin{aligned}\mathbb{E}[Y^4|X]&=\mathbb{E}\left[\sum_{l=0}^4\binom{4}{l}m(X)^l e^{4-l}\middle|X\right]\\&=\sum_{l=0}^4\binom{4}{l}m(X)^l\mathbb{E}e^{4-l}<C'.\end{aligned}$$

This implies that $\mathbb{E}_{Y|X}||Y||_1^4=O(n)$. Combining this with equation (8) and writing $\mathbb{E}_X\left[\sigma_{\min}^+(\mathbb{X})^{-4}\right]=O(n^{-u})$, with $u\geq 0$, we see that

$$\mathbb{E}\left[\left|\left|\widehat{\beta}^0\right|\right|_1^4\right]=O(n^{-u+2\alpha+1}).$$

Therefore, Condition 1 follows if $n^{-u+2\alpha+1}=o\left(\frac{n}{\log n}\right)$, which happens if, for instance, $u>2\alpha$. So, if with high probability $\sigma_{\min}^+(\mathbb{X})=\Omega(n^{\alpha/2})$, i.e. is

of larger order than $n^{\alpha/2}$, Condition 1 holds. Indeed, as shown by Rudelson & Vershynin (2009), a random matrix composed of independent and identically distributed sub-Gaussian random variables has, with high probability,

$$\sigma^{+}_{\min}(\mathbb{X}) \geq \begin{cases} \sqrt{p_n} = n^{\alpha/2} & \text{if } \alpha > 1 \\ \sqrt{n} = n^{1/2} & \text{if } \alpha < 1. \end{cases}$$

In either case, $\sigma^{+}_{\min}(\mathbb{X})$ is at least of order $n^{\alpha/2}$.

4. Proof of main results

To show the results of this paper, we decompose the excess risk into several parts. Define $t_* := \min\{t_{\max}, t_n\}$. Then, we write

$$\begin{aligned} \mathcal{E}(\widehat{t}, t_n) &= L\left(\widehat{\beta}_{\widehat{t}}\right) - L\left(\beta_{t_n}\right) \\ &= \underbrace{L\left(\widehat{\beta}_{\widehat{t}}\right) - \widehat{L}_{V_n}\left(\widehat{t}\right)}_{(I)} + \underbrace{\widehat{L}_{V_n}\left(\widehat{t}\right) - \widehat{L}_{V_n}\left(t_*\right)}_{(II)} \\ &+ \underbrace{\widehat{L}_{V_n}\left(t_*\right) - \widehat{L}\left(\widehat{\beta}_{t_*}\right)}_{(III)} + \underbrace{\widehat{L}\left(\widehat{\beta}_{t_*}\right) - \widehat{L}\left(\widehat{\beta}_{t_n}\right)}_{(IV)} \\ &+ \underbrace{\widehat{L}\left(\widehat{\beta}_{t_n}\right) - L\left(\widehat{\beta}_{t_n}\right)}_{(V)} + \underbrace{L\left(\widehat{\beta}_{t_n}\right) - L\left(\beta_{t_n}\right)}_{(VI)}. \end{aligned}$$

For any $t \in T_n$, $\widehat{L}_{V_n}\left(\widehat{t}\right) - \widehat{L}_{V_n}\left(t\right) \leq 0$. Therefore, as $t_* \in T_n$, $(II) \leq 0$. Also, by the discussion in Section 2.2,

$$\widehat{L}\left(\widehat{\beta}_t\right) = \begin{cases} \widehat{L}\left(\widehat{\beta}_t\right) & \text{if } t < t_{\max} \\ \widehat{L}\left(\widehat{\beta}_{t_{\max}}\right) & \text{if } t \geq t_{\max}. \end{cases}$$

To see this, note that for any $t \geq t_{\max}$, $\widehat{\beta}_t$ is a least squares solution. Therefore, by the definition of t_*, $\widehat{L}\left(\widehat{\beta}_{t_*}\right) = \widehat{L}\left(\widehat{\beta}_{t_n}\right)$ and hence $(IV) = 0$.

To bound the remaining terms, we rewrite them as quadratic forms (Section 4.1) and present three lemmas (Section 4.2). The actual proofs are contained in Section 4.3.

4.1. Squared-error loss and quadratic forms

We can rewrite the notation from Section 2 as quadratic forms. Define the parameter to be $\gamma^\top := (-1, \beta^\top)$, with associated estimator $\widehat{\gamma}_t^\top := (-1, \widehat{\beta}_t^\top)$. We can rewrite equation (2) as

$$L\left(\beta\right) = \mathbb{E}_{F_n}\left[\left(\mathcal{Y} - \mathcal{X}^\top\beta\right)^2\right] = \gamma^\top \Sigma_n \gamma \tag{9}$$

where $\Sigma_n := \mathbb{E}_{F_n}[\mathcal{Z}\mathcal{Z}^\top]$. Analogously, equation (3) has the following form

$$\widehat{L}\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top\beta)^2 = \gamma^\top\widehat{\Sigma}_n\gamma,$$

where $\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n Z_i Z_i^\top$. Lastly, we rewrite equation (4) as

$$\widehat{L}_{V_n}\left(t\right) = \frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_t^{(v)})^T\widehat{\Sigma}_v\widehat{\gamma}_t^{(v)}, \tag{10}$$

where $\widehat{\Sigma}_v = |v|^{-1}\sum_{r\in v} Z_r Z_r^\top$, $(\widehat{\gamma}_t^{(v)})^\top := (-1, (\widehat{\beta}_t^{(v)})^\top)$, and

$$\widehat{\beta}_t^{(v)} := \operatorname*{argmin}_{\beta\in\mathcal{B}_t}\gamma^\top\widehat{\Sigma}_{(v)}\gamma,$$

with $\widehat{\Sigma}_{(v)} := (n - |v|)^{-1}\sum_{r\notin v} Z_r Z_r^\top$.

With this notation, each part of the decomposition can be written as the difference of quadratic forms. Careful modifications will allow us to use the following lemmas to prove bounds for each part.

4.2. Supporting lemmas

Several times in our proof of the main results we need to bound a quadratic form given by a symmetric matrix and an estimator indexed by a tuning parameter. To this end, we state and prove the following simple lemma.

Lemma 4.1. *Suppose $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$. Then*

$$a^\top A a \leq ||a||_1^2\,|||A|||_\infty,$$

where $|||A|||_\infty := \max_{i,j}|A_{ij}|$ is the entry-wise max norm.

Proof of Lemma 4.1.

$$\begin{aligned} a^\top A a &\leq ||a||_1\,||Aa||_\infty \\ &\leq ||a||_1 \max_{ij}\{|A_{ij}|\}\,||a||_1 \\ &= ||a||_1^2\,|||A|||_\infty, \end{aligned}$$

where the first inequality follows by Hölder's inequality. □

Additionally, we include Nemirovski's inequality for completeness. See Nemirovski (2000) or Dümbgen et al. (2010) for details.

Lemma 4.2 (Nemirovski's inequality)**.** *Let $\xi_1, \ldots, \xi_n$ be independent random vectors in $\mathbb{R}^d$, for $d \geq 3$ with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[||\xi_i||_2^2] < \infty$. Then for every*

$s \in [2, \infty]$, and index set v, there exists an absolute constant $\widetilde{C}$ (independent of s, n, d, v, and the distribution of the ξ_i's) such that

$$\mathbb{E}\left[\left|\left|\sum_{i\in v}\xi_i\right|\right|_s^2\right] \leq \widetilde{C}\min(s,\log d)\sum_{i\in v}\mathbb{E}\left[||\xi_i||_s^2\right],$$

where $||\cdot||_s$ is the ℓ_s norm.

Finally, we will use Lemma 4.2 to find the rate of convergence for the sample covariance matrix to the population covariance.

Lemma 4.3. *Let $V_n = \{v_1, \ldots, v_{K_n}\}$ be a set of validation sets satisfying Condition 2. Then,*

$$\mathbb{E}\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty\right)^2 = O\left(\frac{\log n}{c_n}\right).$$

Proof of Lemma 4.3. First, note that

$$\begin{aligned}
&\mathbb{E}\left(\sum_{v\in V_n}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty\right)^2 \\
&= \sum_{u\neq v\in V_n}\mathbb{E}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty \mathbb{E}\left|\left|\left|\widehat{\Sigma}_u - \Sigma_n\right|\right|\right|_\infty + \\
&\quad + \sum_{v\in V_n}\mathbb{E}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty^2
\end{aligned}$$

by independence and disjoint elements of v. Let $\xi_r \in \mathbb{R}^{(p+1)^2}$ be the vectorized version of the zero-mean matrix $\frac{1}{c_n}(Z_r Z_r^\top - \mathbb{E}\mathcal{Z}\mathcal{Z}^\top)$. Then,

$$\begin{aligned}
\mathbb{E}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty &\leq \sqrt{\mathbb{E}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty^2} \\
&= \sqrt{\mathbb{E}\left|\left|\sum_{r\in v}\xi_r\right|\right|_\infty^2}
\end{aligned}$$

by Jensen's inequality. Using Lemma 4.2 with $s = \infty$ and $d = (p+1)^2$, we find

$$\begin{aligned}
&\mathbb{E}\left|\left|\sum_{r\in v}\xi_r\right|\right|_\infty^2 \\
&\leq \widetilde{C}\log\left((p+1)^2\right)\sum_{r\in v}\mathbb{E}\,||\xi_r||_\infty^2 \\
&\lesssim \log(4n^{2\alpha})\frac{1}{c_n^2}\sum_{r\in v}\mathbb{E}\left[\max_{0\leq j,k\leq p}|Z_{rj}Z_{rk} - \mathbb{E}\mathcal{Z}_j\mathcal{Z}_k|^2\right] \\
&\leq \log(4n^{2\alpha})\frac{1}{c_n^2}\sum_{r\in v}C \lesssim \frac{\log n}{c_n}.
\end{aligned}$$

where second to last inequality follows by Definition 3.1 with associated constant C. Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|\right|_\infty\right)^2 \\
&\lesssim \frac{1}{K_n^2}\left(\sum_{u,v\in V_n}\frac{\log n}{c_n}\right) = \frac{\log n}{c_n}.
\end{aligned}$$

□

4.3. Proof of theorems

We break this section into parts based on the decomposition of equation (6).

Final predictor and cross-validation risk (I) Note that by equation (9) and equation (10)

$$\begin{aligned}
&L\left(\widehat{\beta}_{\widehat{t}}\right) - \widehat{L}_{V_n}\left(\widehat{t}\right) \\
&= \widehat{\gamma}_{\widehat{t}}^\top\,\Sigma_n\,\widehat{\gamma}_{\widehat{t}} - \frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^\top\widehat{\Sigma}_v\widehat{\gamma}_{\widehat{t}}^{(v)} \\
&= \left[\widehat{\gamma}_{\widehat{t}}^\top\,\Sigma_n\,\widehat{\gamma}_{\widehat{t}} - \widehat{\gamma}_{\widehat{t}}^\top\left(\widehat{\Sigma}_n\right)\widehat{\gamma}_{\widehat{t}}\right] + \\
&\quad \left[\widehat{\gamma}_{\widehat{t}}^\top\left(\widehat{\Sigma}_n\right)\widehat{\gamma}_{\widehat{t}} - \frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^\top\widehat{\Sigma}_v\widehat{\gamma}_{\widehat{t}}^{(v)}\right]
\end{aligned}$$

Addressing each of the terms in order,

$$\begin{aligned}
&\left[\widehat{\gamma}_{\widehat{t}}^\top\,\Sigma_n\,\widehat{\gamma}_{\widehat{t}} - \widehat{\gamma}_{\widehat{t}}^\top\left(\widehat{\Sigma}_n\right)\widehat{\gamma}_{\widehat{t}}\right] \\
&= \widehat{\gamma}_{\widehat{t}}^\top\left(\Sigma_n - \widehat{\Sigma}_n\right)\widehat{\gamma}_{\widehat{t}} \\
&\leq \sup_{t\in T_n}\sup_{\beta\in\mathcal{B}_t}||\gamma_t||_1^2\left|\left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|\right|_\infty \\
&\leq \left|\left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|\right|_\infty \sup_{t\in T_n}(1+t)^2 \\
&= \left|\left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|\right|_\infty (1+t_{\max})^2.
\end{aligned}$$

The first inequality follows by Lemma 4.1 while the second inequality is by the definition of $\mathcal{B}_t$, and the equality in the last line follows by the definition T_n.

Likewise,

$$\begin{aligned}
&\left[\widehat{\gamma}_{\widehat{t}}^{\top}\left(\widehat{\Sigma}_n\right)\widehat{\gamma}_{\widehat{t}}-\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_v\widehat{\gamma}_{\widehat{t}}^{(v)}\right]\\
&=\left(\widehat{\gamma}_{\widehat{t}}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}-\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}^{(v)}\right)+\\
&\quad+\left(\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}^{(v)}-\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_v\widehat{\gamma}_{\widehat{t}}^{(v)}\right)\\
&=\frac{1}{K_n}\sum_{v\in V_n}\left(\widehat{\gamma}_{\widehat{t}}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}-(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}^{(v)}\right)+\\
&\quad+\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\left(\widehat{\Sigma}_n-\widehat{\Sigma}_v\right)\widehat{\gamma}_{\widehat{t}}^{(v)}\\
&\leq\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\left(\widehat{\Sigma}_n-\widehat{\Sigma}_v\right)\widehat{\gamma}_{\widehat{t}}^{(v)}.
\end{aligned}$$

The last inequality follows as $\widehat{\gamma}_{\widehat{t}}$ is chosen to minimize $\widehat{\gamma}_{\widehat{t}}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}$ over $\mathcal{B}_{\widehat{t}}$, and so for any $v\in V_n$, $\widehat{\gamma}_{\widehat{t}}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}\leq(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{\widehat{t}}^{(v)}$.

Continuing and using Lemma 4.1,

$$\begin{aligned}
&\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{\widehat{t}}^{(v)})^{\top}\left(\widehat{\Sigma}_n-\widehat{\Sigma}_v\right)\widehat{\gamma}_{\widehat{t}}^{(v)}\\
&\leq\frac{1}{K_n}\sum_{v\in V_n}\sup_{t\in T_n}\sup_{\beta\in B_t}||\gamma||_2^2\left|\left|\left|\widehat{\Sigma}_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\\
&\leq\frac{1}{K_n}\sum_{v\in V_n}\sup_{t\in T_n}\sup_{\beta\in B_t}||\gamma||_1^2\left|\left|\left|\widehat{\Sigma}_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\\
&\leq(1+t_{\max})^2\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\widehat{\Sigma}_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\\
&\leq(1+t_{\max})^2\left(\frac{1}{K_n}\right)\cdot\\
&\qquad\sum_{v\in V_n}\left(\left|\left|\left|\widehat{\Sigma}_n-\Sigma_n\right|\right|\right|_{\infty}+|||\Sigma_n-\Sigma_{\{r\}}|||_{\infty}\right)\\
&=(1+t_{\max})^2\cdot\\
&\qquad\left(\left|\left|\left|\Sigma_n-\widehat{\Sigma}_n\right|\right|\right|_{\infty}+\frac{1}{K_n}\sum_{v\in V_n}|||\Sigma_n-\Sigma_{\{r\}}|||_{\infty}\right).
\end{aligned}$$

Combining these results together, we obtain the following upper bound for (I)

$$\begin{aligned}
&L\left(\widehat{\beta}_{\widehat{t}}\right)-\widehat{L}_{V_n}\left(\widehat{t}\right)\\
&\leq(1+t_{\max})^2\cdot\\
&\qquad\left(2\left|\left|\left|\Sigma_n-\widehat{\Sigma}_n\right|\right|\right|_{\infty}+\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\Sigma_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\right).
\end{aligned}$$

By Lemma 4.3 with $V_n=\{\{1,\ldots,n\}\}$ and $c_n=n$,

$$\mathbb{E}\left|\left|\left|\Sigma_n-\widehat{\Sigma}_n\right|\right|\right|_{\infty}^2=O\left(\frac{\log n}{n}\right).$$

Additionally, by Lemma 4.3 with $V_n=\{v_1,\ldots,v_{K_n}\}$,

$$\mathbb{E}\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\Sigma_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\right)^2=O\left(\frac{\log n}{c_n}\right).$$

Furthermore, $\mathbb{E}\left[(1+t_{\max})^4\right]\asymp\mathbb{E}\left[(t_{\max})^4\right]$ implies $\mathbb{E}\left[(1+t_{\max})^4\right]\asymp(t_n^4)$. Combining these three bounds together, we get

$$\begin{aligned}
&\mathbb{E}|(I)|\\
&\leq\sqrt{\mathbb{E}\left[(1+t_{\max})^4\right]\mathbb{E}\left|\left|\left|\Sigma_n-\widehat{\Sigma}_n\right|\right|\right|_{\infty}^2}+\\
&\quad+\sqrt{\mathbb{E}\left[(1+t_{\max})^4\right]\mathbb{E}\left[\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\Sigma_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\right)^2\right]}\\
&=o\left(\sqrt{t_n^4\frac{\log n}{n}}\right)+o\left(\sqrt{t_n^4\frac{\log n}{c_n}}\right)=o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right).
\end{aligned}$$

Hence,

$$\mathbb{E}\left(L\left(\widehat{\beta}_{\widehat{t}}\right)-\widehat{L}_{V_n}\left(\widehat{t}\right)\right)=o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right)$$

for any V_n.

Cross-validation risk and empirical risk (III) Recall that $\widehat{\Sigma}_{(v)}=\frac{1}{n-c_n}\sum_{r\notin v}Z_rZ_r^{\top}$.

Then,

$$\begin{aligned}
&\widehat{L}_{V_n}(t_*)-\widehat{L}\left(\widehat{\beta}_{t_*}\right)\\
&=\frac{1}{K_n}\sum_{v\in V_n}(\widehat{\gamma}_{t_*}^{(v)})^{\top}\widehat{\Sigma}_v\widehat{\gamma}_{t_*}^{(v)}-\widehat{\gamma}_{t_*}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{t_*}\\
&=\frac{1}{K_n}\sum_{v\in V_n}\left((\widehat{\gamma}_{t_*}^{(v)})^{\top}\widehat{\Sigma}_v\widehat{\gamma}_{t_*}^{(v)}-(\widehat{\gamma}_{t_*}^{(v)})^{\top}\widehat{\Sigma}_{(v)}\widehat{\gamma}_{t_*}^{(v)}\right)+\\
&\quad+\frac{1}{K_n}\sum_{v\in V_n}\left((\widehat{\gamma}_{t_*}^{(v)})^{\top}\widehat{\Sigma}_{(v)}\widehat{\gamma}_{t_*}^{(v)}-\widehat{\gamma}_{t_*}^{\top}\widehat{\Sigma}_n\widehat{\gamma}_{t_*}\right)\\
&\leq\frac{1}{K_n}\sum_{v\in V_n}(1+t_*)^2\left|\left|\left|\widehat{\Sigma}_v-\widehat{\Sigma}_{(v)}\right|\right|\right|_{\infty}\\
&\leq(1+t_*)^2\cdot\\
&\qquad\frac{1}{K_n}\sum_{v\in V_n}\left(\left|\left|\left|\Sigma_n-\widehat{\Sigma}_{(v)}\right|\right|\right|_{\infty}+\left|\left|\left|\Sigma_n-\widehat{\Sigma}_v\right|\right|\right|_{\infty}\right).
\end{aligned}$$

The second-to-last inequality follows by Lemma 4.1 and the fact that $\widehat{\gamma}_{t_*}^{(v)}$ is chosen to minimize

$(\widehat{\gamma}_{t_*}^{(v)})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_*}^{(v)}$, which implies $(\widehat{\gamma}_{t_*}^{(v)})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_*}^{(v)} \leq \widehat{\gamma}_{t_*}^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_*}$.

Using a straight-forward adaptation of Lemma 4.3

$$\mathbb{E}\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\Sigma_n - \widehat{\Sigma}_{(v)}\right|\right|\right|_\infty\right)^2 = O\left(\frac{\log n}{n-c_n}\right).$$

As $n > c_n$, by assumption, we see

$$\mathbb{E}\left(\frac{1}{K_n}\sum_{v\in V_n}\left|\left|\left|\Sigma_n - \widehat{\Sigma}_{(v)}\right|\right|\right|_\infty\right)^2 = O\left(\frac{\log n}{n}\right).$$

Therefore, following the analogous steps established in the proof of (I), we get

$$\mathbb{E}\left(\widehat{L}_{V_n}(t_*) - \widehat{L}\left(\widehat{\beta}_{t_*}\right)\right) = o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right).$$

Empirical risk and expected risk (V, VI) The proof of these results follows from the results established in Greenshtein & Ritov (2004). We include a somewhat different proof for completeness. Observe the following bounds

$$L\left(\widehat{\beta}_{t_n}\right) - \widehat{L}\left(\widehat{\beta}_{t_n}\right) \leq \sup_{\beta\in\mathcal{B}_{t_n}}\left|L(\beta) - \widehat{L}(\beta)\right|$$

and

$$\begin{aligned}
&L\left(\widehat{\beta}_{t_n}\right) - L(\beta_{t_n}) \\
&= L\left(\widehat{\beta}_{t_n}\right) - \widehat{L}\left(\widehat{\beta}_{t_n}\right) + \widehat{L}\left(\widehat{\beta}_{t_n}\right) - L(\beta_{t_n}) \\
&\leq 2\sup_{\beta\in\mathcal{B}_{t_n}}\left|L(\beta) - \widehat{L}(\beta)\right|.
\end{aligned}$$

Therefore, both (V) and (VI) follow since

$$\begin{aligned}
&\mathbb{E}\sup_{\beta\in\mathcal{B}_{t_n}}\left|L(\beta) - \widehat{L}(\beta)\right| \\
&= \mathbb{E}\sup_{\beta\in\mathcal{B}_{t_n}}\left|\gamma^\top\Sigma_n\gamma - \gamma^\top\widehat{\Sigma}_n\gamma\right| \\
&= \mathbb{E}\sup_{\beta\in\mathcal{B}_{t_n}}\left|\gamma^\top(\Sigma_n - \widehat{\Sigma}_n)\gamma\right| \\
&\leq \mathbb{E}\sup_{\beta\in\mathcal{B}(t_n)}(1+||\beta||_1)^2\left|\left|\left|\widehat{\Sigma}_n - \Sigma_n\right|\right|\right|_\infty \\
&\leq (1+t_n)^2\mathbb{E}\left|\left|\left|\widehat{\Sigma}_n - \Sigma_n\right|\right|\right|_\infty = o\left(t_n^2\sqrt{\frac{\log n}{n}}\right)
\end{aligned}$$

by Lemma 4.3.

This completes the proof of Theorem 3.2. In particular, we have shown that

$$\begin{aligned}
&\mathbb{P}\left(L\left(\widehat{\beta}_{\widehat{t}}\right) - L(\beta_{t_n}) > \delta\right) \\
&\leq o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right) + o\left(t_n^2\sqrt{\frac{\log n}{n}}\right) \\
&= o\left(t_n^2\sqrt{\frac{\log n}{c_n}}\right).
\end{aligned}$$

5. Conclusion

A common practice in data analysis is to estimate the coefficients of a linear model via lasso and choose the regularization parameter via cross-validation. Unfortunately, no theoretical results existed as to the effect of choosing the tuning parameter in this data-dependent way.

In this paper, we demonstrate that the lasso with tuning parameter chosen by cross-validation is persistent. This is the first step in a establishing a broader understanding of the interaction between the lasso and a data-dependent tuning parameter. In particular, by imposing an eigenvalue type condition on the design matrix, we can achieve the same risk-consistency results with a data-dependent tuning parameter as with the optimal tuning parameter. We feel that this paper provides some theoretical justification for the received wisdom that cross-validation is a useful tool for the applied researcher in the context of the lasso.

Acknowledgements

The authors would like the thank Cosma Shalizi and Ryan Tibshirani for reading preliminary versions of this manuscript and offering helpful suggestions. We also thank three anonymous referees for their careful and insightful comments and Larry Wasserman for the inspiration.

References

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

Chen, S.S., Donoho, D.L., and Saunders, M.A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Donoho, D.L., Elad, M., and Temlyakov, V.N. Stable recovery of sparse overcomplete representations in the

presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

Dümbgen, L., Van De Geer, S.A., Veraar, M.C., and Wellner, J.A. Nemirovski's inequalities revisited. *American Mathematical Monthly*, 117(2):138–160, 2010.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics*, 32(2): 407–499, 2004.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Fu, W. and Knight, K. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

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

Györfi, L., Kohler, M., Krzyżak, A., and H., Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, 2002.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, 2009.

Leng, C., Lin, Y., and Wahba, G. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

Nemirovski, A. Topics in non-parametric statistics. lectures on probability theory and statistics (Saint-Flour, 1998), 85–277. *Lecture Notes in Math*, 1738:86–277, 2000.

Osborne, M.R., Presnell, B., and Turlach, B.A. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

Plutowski, M., Sakata, S., and White, H. Cross-validation estimates IMSE. *Advances in Neural Information Processing Systems*, 6, 1994.

Rudelson, M. and Vershynin, R. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

Shao, J. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Tibshirani, R.J. and Taylor, J. Degrees of freedom in lasso problems. *Annals of Statistics*, 40:1198–1232, 2012.

van de Geer, S. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

van de Geer, S. and Lederer, J. The lasso, correlated design, and improved oracle inequalities, 2011. URL http://arxiv.org/abs/1107.0189.

Wainwright, M.J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Wang, H. and Leng, C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

Xu, H., Mannor, S., and Caramanis, C. Sparse algorithms are not stable: A no-free-lunch theorem. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1299–1303. IEEE, 2008.

Zhao, P. and Yu, B. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7: 2541–2563, 2006.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429, 2006.

Zou, H., Hastie, T., and Tibshirani, R. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5): 2173–2192, 2007.