Statistica Sinica **27** (2017), 1017-1036
doi:https://doi.org/10.5705/ss.202015.0355

RISK CONSISTENCY OF CROSS-VALIDATION WITH LASSO-TYPE PROCEDURES

Darren Homrighausen and Daniel J. McDonald

Colorado State University and Indiana University

Abstract: The lasso and related sparsity inducing algorithms have been the target of substantial theoretical and applied research. Correspondingly, many results are known about their behavior for a fixed or optimally chosen tuning parameter specified up to unknown constants. In practice, however, this oracle tuning parameter is inaccessible so one must use the data to select one. Common statistical practice is to use a variant of cross-validation for this task. However, little is known about the theoretical properties of the resulting predictions with such data-dependent methods. We consider the high-dimensional setting with random design wherein the number of predictors p grows with the number of observations n. Under typical assumptions on the data generating process, similar to those in the literature, we recover oracle rates up to a log factor when choosing the tuning parameter with cross-validation. Under weaker conditions, when the true model is not necessarily linear, we show that the lasso remains risk consistent relative to its linear oracle. We also generalize these results to the group lasso and square-root lasso and investigate the predictive and model selection performance of cross-validation via simulation.

Key words and phrases: Linear oracle, model selection, persistence, regularization.

1. Introduction

Following its introduction in the statistical (Tibshirani (1996)) and signal processing (Chen, Donoho and Saunders (1998)) communities, ℓ_1-regularized linear regression has become a fixture as both a data analysis tool and as a subject for 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 λ, we consider the lasso problem of finding

$$\widehat{\beta}_\lambda = \operatorname*{argmin}_{\beta} \frac{1}{n} \left|\left|Y - \mathbb{X}\beta\right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1 \tag{1.1}$$

for any λ, where $||\cdot||_2$ and $||\cdot||_1$ indicate the Euclidean and ℓ_1-norms, respectively. An equivalent but less computationally convenient specification of the lasso problem given by (1.1) is the constrained optimization version:

$$\widehat{\beta}_t := \widehat{\beta}(\mathcal{B}_t) \in \operatorname*{argmin}_{\beta\in\mathcal{B}_t} \frac{1}{n}\,||Y - \mathbb{X}\beta||_2^2\,, \tag{1.2}$$

where $\mathcal{B}_t := \{\beta : ||\beta||_1 \leq t\}$. By convexity, for each λ (or t), there is always at least one solution to these optimization problems. While it is true that the solution is not necessarily unique if $\text{rank}(\mathbb{X}) < p$, this detail is unimportant for our purposes, and we abuse notation slightly by referring to $\widehat{\beta}_\lambda$ as 'the' lasso solution.

These two optimization problems are equivalent mathematically, but they differ substantially in practice. Though the constrained optimization problem, (1.2) can be solved via quadratic programming, most algorithms use the Lagrangian formulation (1.1). In this paper, we address both estimators as each is more amenable to theoretical treatment in different contexts.

The literature contains numerous results regarding the statistical properties of the lasso, and, while it is beyond the scope of this paper to give a complete literature review, we highlight some of these results here. Early results about the asymptotic distribution of the lasso solution are shown in Knight and Fu (2000) under the assumption that the sample covariance matrix has a nonnegative definite limit and p is fixed. Many authors (e.g. Donoho, Elad and Temlyakov (2006); Meinshausen and Bühlmann (2006); Meinshausen and Yu (2009); Wainwright (2009); Zhao and Yu (2006); Zou (2006)) have investigated model selection properties of the lasso—showing that when the best predicting model is linear and sparse, the lasso will tend to asymptotically recover those predictors. The literature has considered this setting under various "irrepresentability" conditions which ensure that the relevant predictors are not too highly correlated with irrelevant ones. Bickel, Ritov and Tsybakov (2009) analyze the lasso and the Dantzig selector Candès and Tao (2007) under restricted eigenvalue conditions with an oracle tuning parameter. Finally, Belloni, Chernozhukov and Wang (2014) develop results for a related method, the square-root lasso, with heteroscedastic noise and oracle tuning parameter.

Theoretical results such as these and others, depend critically on the choice of tuning parameters and are typically of the form: if $t = t_n = o(n/\log p)^{1/4}$, then $\widehat{\beta}_{t_n}$ possesses some desired behavior (correct model selection, risk consistency, sign consistency, et cetera). 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 way.

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 matrix $\mathbb{X}$ and the response vector Y. Additionally, we demonstrate that our framework can be used to show similar results for other lasso-type methods. Our results build on the previous theory presented in Homrighausen and McDonald (2014) and Homrighausen and McDonald (2013). Homrighausen and McDonald (2014) proves risk consistency for cross-validation under strong conditions on the data generating process, most notably $n > p$, and on the cross-validation procedure (requiring leave-one-out CV). The results in this paper differ from those in Homrighausen and McDonald (2013) in a number of ways. The current paper examines the Lagrangian formulation of the lasso problem under typical conditions, weakens the conditions on an upper bound for t, provides more refined results via concentration inequalities, examines the influence of K, and includes related results for the group lasso and the square-root lasso.

1.1. Overview of results

We focus on risk consistency, (alternatively known as persistence), investigating the difference between the prediction risk of the lasso estimator with tuning parameter estimated by cross-validation and the risk of the best linear oracle predictor (with oracle tuning parameter). Risk consistency of lasso has previously been studied by Greenshtein and Ritov (2004); Bunea, Tsybakov and Wegkamp (2007); van de Geer (2008); Bartlett, Mendelson and Neeman (2012). Their results, in contrast to ours, assume that the tuning parameter is selected in an oracle fashion.

We present two results that make progressively stronger assumptions on the data generating process and use both forms of the lasso estimator. The first imposes strong conditions on the design matrix, assumes the linear model is true, and that this linear model is sparse. The second allows the true model to be neither linear nor correctly specified. Our focus is on risk consistency rather than estimation of a “true” parameter or correct identification of a “true” sparsity pattern. Additionally, well-known results of Shao (1993) imply that cross-validation leads to inconsistent model selection in general, suggesting that results for sparse recovery may not exist. Although prediction is an important goal, one is often interested in variable selection for more interpretable models or follow-up experiments. In light of the negative results in Shao (1993), we are unable to offer theoretical results that promise consistent model selection by cross validation, but simulations in Section 4 suggest that it performs well nevertheless. Both the estimation and sparse recovery settings are frequently studiedassuming the tuning parameter is the oracle and that the data generating model is linear (e.g. Bunea, Tsybakov and Wegkamp (2007); Candès and Plan (2009); Donoho, Elad and Temlyakov (2006); Leng, Lin and Wahba (2006); Meinshausen and Bühlmann (2006); Meinshausen and Yu (2009)).

In our first case, when the truth is linear, we examine the Lagrangian formulation in (1.1). We prove convergence rates which differ only by a log factor relative to those achievable with the oracle tuning parameter (e.g. Negahban et al. (2012); Bühlmann and van de Geer (2011); Bunea, Tsybakov and Wegkamp (2007)). Thus for an s^*-sparse linear model with restricted isometry conditions on the covariance of the design, the risk of the cross-validated estimator approaches the oracle risk at a rate of $O(s^* \log(p)\log(n)/n)$. Under more general conditions, we follow the approach of Greenshtein and Ritov (2004) and examine the constrained optimization form in (1.2). Using our methods, we require that the set of candidate predictors, $\mathcal{B}_{t_n}$, satisfies $t_n = o\left(n^{1/4}/(m_n(\log p)^{1/4+1/(2q))}\right)$ where m_n is a sequence that approaches infinity and q characterizes the tail behavior of the data. This is essentially as quickly as one could hope relative to Greenshtein and Ritov (2004) under our more general assumptions on the design matrix. We note however that, using empirical process techniques, Bartlett, Mendelson and Neeman (2012) have been able to improve the rate shown in Greenshtein and Ritov (2004) to $t_n = o(n^{1/2}/(\log^{3/2} n \log^{3/2}(np)))$ for sub-Gaussian design and an oracle tuning parameter.

1.2. Tuning parameter selection methods and outline of the paper

There are several proposed data-dependent techniques for choosing t or λ. Kato (2009) and Tibshirani and Taylor (2012) investigate estimating the "degrees of freedom" of a lasso solution. With an unbiased estimator of the degrees of freedom, the tuning parameter can be selected by minimizing the empirical risk penalized by a function of this estimator. This approach requires an estimate of the variance, which is nontrivial when $p > n$ Giraud, Huet and Verzelen (2012). Another risk estimator is the adapted Bayesian information criterion proposed by Wang and Leng (2007) that uses a plug-in estimator based on the second-order Taylor's expansion of the risk. Arlot and Massart (2009) and Saumard (2011) consider "slope heuristics" as a method for penalty selection with Gaussian noise, paying particular attention to the regressogram estimator in the first case and piecewise polynomials with p fixed in the second. Sun and Zhang (2012) present an algorithm to jointly estimate the regression coefficients and the noise level; this results in a data-driven value for the tuning parameter that possesses oracle properties under some regularity conditions.

Many authors (e.g Efron et al. (2004); Friedman, Hastie and Tibshirani (2010); Greenshtein and Ritov (2004); Tibshirani (1996, 2011); Zou, Hastie and Tibshirani (2007)) and as discussed by Bühlmann and van de Geer (2011), Sec.2.4.1 recommend selecting t or λ in the lasso problem by minimizing a K-fold cross-validation estimator of the risk (see Section 2 for the precise definition). Cross-validation is generally well-studied in a number of contexts, especially model selection and risk estimation. In the context of model selection, Arlot and Celisse (2010) give a detailed survey of the literature emphasizing the relationship between the sizes of the validation set and the training set, as well as discussing the positive bias of cross-validation as a risk estimator.

Some results supporting the use of cross-validation for statistical methods other than lasso are known. For instance, Stone (1974, 1977) outlines various conditions under which cross-validated methods can result in good predictions. Dudoit and van der Laan (2005) find finite sample bounds for various cross-validation procedures. These results do not address the lasso nor parameter spaces with increasing dimensions and, furthermore, apply to choosing the best estimator from a finite collection of candidate estimators. Lecué and Mitchell (2012) provide oracle inequalities related to using cross-validation with lasso, however, they treat the problem as aggregating a dictionary of lasso estimators with different tuning parameters, and the results are stated for fixed p rather than the high-dimensional setting investigated here. Flynn, Hurvich and Simonoff (2013) investigate numerous methods for tuning parameter selection in penalized regression, but the theoretical results hold only when $p/n \to 0$ and not for cross-validation. In particular, the authors state "to our knowledge the asymptotic properties of $[K]$-fold CV have not been fully established in the context of penalized regression" ((Flynn, Hurvich and Simonoff, 2013, p.1,033)).

Rather than cross-validation, one may use information criteria such as AIC Akaike (1974) or BIC Schwarz (1978). These techniques are frequently advocated (e.g. Bühlmann and van de Geer (2011); Wang, Li and Tsai (2007); Tibshirani (1996); Fan and Li (2001)), but the classical asymptotic arguments underlying these methods apply only for p fixed and rely on maximum likelihood estimates (or Bayesian posteriors) for all parameters including the noise. The theory in the high-dimensional setting supporting these methods is less complete. Recent work has developed new information criteria with supporting asymptotic results if $\text{rank}(\mathbb{X}) = p$ but is still allowed to increase. For example, the criterion developed by Wang, Li and Leng (2009) selects the correct model asymptotically even if

$p \to \infty$ as long as $p/n \to 0$. If p is allowed to increase more quickly than n, theoretical results assume σ^2 is known to get around the difficult task of high-dimensional variance estimation (Chen and Chen (2012); Zhang and Shen (2010); Kim, Kwon and Choi (2012); Fan and Tang (2013)).

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 compares different choices of K for cross-validation via simulation, while Section 5 presents some avenues for further research.

2. Notation and Definitions

2.1. Preliminaries

Suppose we observe data $Z_i^\top = (Y_i, X_i^\top)$ consisting of predictor variables, $X_i \in \mathbb{R}^{p_n}$, and response variables, $Y_i \in \mathbb{R}$, where $Z_i \sim \mu_n$ are independent and identically distributed for $i = 1, 2, \ldots, n$ and the distribution μ_n is in some class $\mathcal{F}$ to be specified. Here, we use the notation p_n to allow the number of predictor variables to change with n. Similarly, we index the distribution μ_n to emphasize its dependence on n. For simplicity, we omit the subscript n when there is little risk of confusion.

We consider the problem of estimating a linear functional $f(\mathcal{X}) = \mathcal{X}^\top\beta$ for predicting $\mathcal{Y}$, when $\mathcal{Z}^\top = (\mathcal{Y}, \mathcal{X}^\top) \sim \mu_n$ is a new, independent random variable from the same distribution as the data and $\beta = (\beta_1, \ldots, \beta_p)^\top$. For now, we assume only the usual regression framework where $\mathcal{Y} = f^*(\mathcal{X}) + \epsilon$, with ϵ and $\mathcal{X}$ independent and f^* is some unknown function. We use zero-based indexing for $\mathcal{Z}$ so that $\mathcal{Z}_0 = \mathcal{Y}$. To measure performance, we use the L^2-risk of the predictor β:

$$R(\beta) := \mathbb{E}_{\mu_n}\left[(\mathcal{Y} - \mathcal{X}^\top\beta)^2\right]. \tag{2.1}$$

This is a conditional expectation: for any estimator $\widehat{\beta}$,

$$R\left(\widehat{\beta}\right) := \mathbb{E}_{\mu_n}\left[(\mathcal{Y} - \mathcal{X}^\top\widehat{\beta})^2 | Z_1, \ldots, Z_n\right], \tag{2.2}$$

and the expectation is taken only over the new data $\mathcal{Z}$ and not over any observables which may be used to choose $\widehat{\beta}$.

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

$$\widehat{R}\,(\beta) := \frac{1}{n}||Y - \mathbb{X}\beta||_2^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2. \tag{2.3}$$

Analogously to (2.3), we write the *K-fold cross-validation estimator of the risk with respect to the tuning parameter t*, abbreviated as CV-risk or just CV, as

$$\widehat{R}_{V_n}\,(t) = \widehat{R}_{V_n}\left(\widehat{\beta}_t^{(v_1)}, \ldots, \widehat{\beta}_t^{(v_K)}\right) := \frac{1}{K}\sum_{v \in V_n}\frac{1}{|v|}\sum_{r \in v}\left(Y_r - X_r^\top \widehat{\beta}_t^{(v)}\right)^2. \tag{2.4}$$

Here, $V_n = \{v_1, \ldots, v_K\}$ is a set of validation sets, $\widehat{\beta}_t^{(v)}$ is the estimator in (1.2) with the observations in the validation set v removed, and $|v|$ indicates the cardinality of v. In particular, the cross-validation estimator of the risk is a function of t rather than a single predictor β—it uses K different predictors at a fixed t, averaging over their performance on the respective held-out sets. Over the course of the paper, we freely exchange λ for t in this definition.

The CV-risk minimizing choice of tuning parameter is

$$\widehat{t} = \operatorname*{argmin}_{t \in T} \widehat{R}_{V_n}\,(t)\,. \tag{2.5}$$

In our setting, we take T (or Λ) as an interval subset of the nonnegative real numbers that needs to be defined by the data-analyst. The choice of T is an important part of the performance of $\widehat{\beta}_{\widehat{t}}$ and requires some explanation. First, we provide some insight into the computational load of using CV-risk to find $\widehat{t}$.

2.2. Computations

CV-risk is known to be time consuming and somewhat unstable due to the randomness associated with forming V_n. For a fixed $v \in V_n$, suppose $\widehat{\beta}_\lambda^{(v)}$ is found for the entire lasso path via the `lars` Efron et al. (2004) algorithm, which can be computed in the same computational complexity as a least squares fit. To fix ideas, suppose $n > p$, which means `lars` has computational complexity $O(np^2 + p^3)$. Hence, as the feature matrix $\mathbb{X}$ with $|v|$ rows removed has approximately $n(K-1)/K$ rows, $\widehat{\beta}_\lambda^{(v)}$ can be computed for all λ in $O((n(K-1)/K)p^2 + p^3)$ time. Repeating this K times means the computational complexity for forming $\widehat{R}_{V_n}\,(\lambda)$ over all λ is $O(n(K-1)p^2 + p^3)$. If K is a fixed fraction of n, CV-risk has computational complexity of order $(np)^2$, which is a factor of n slower than a single lasso fit.

Though more expensive on a single processor, CV-risk is readily parallelizable over the K folds and therefore (ignoring communication costs between processors) CV-risk could actually be *faster* to compute than $\widehat{R}$ (and hence $\widehat{\beta}_\lambda$) as $n(K-1)/K < n$. This advantage is lost if we ultimately compute $\widehat{\beta}_{\widehat{\lambda}}$. However, this computational advantage would be maintained if we instead report

$$\tilde{\beta} = \frac{1}{K} \sum_{v \in V_n} \tilde{\beta}^{(v)}, \tag{2.6}$$

where $\tilde{\beta}^{(v)}$ is the lasso estimate trained on the observations in (v) with the tuning parameter chosen by minimizing the empirical risk using the test observations in v. For $K = 4$, for example, this provides a 25% reduction in computation time. The properties of this approximation is an interesting avenue for further investigation.

2.3. Choosing the sets Λ and T

The data analyst must be able to solve the optimization problem in Equation (2.5). For Λ, we must choose a lower bound: $\Lambda = [\lambda_n, \infty)$. This implies we must choose λ_n as a function of the data. While it is tempting to allow $\lambda_n = 0$, this results in numerical and practical implementation issues if $\operatorname{rank}(\mathbb{X}) < p$ and is unnecessary as the theory will show. However, the lower bound has a nontrivial impact on the quality of the recovery, as choosing a value too large may eliminate the best solutions. We see that, under some assumptions on the data generating process, one can safely choose a particular $\lambda_n > 0$ that allows order $\log n$ coefficients to be selected without compromising the quality of the estimator.

In the case of T, an upper bound must be selected for any grid-search optimization procedure. As we impose much weaker conditions on the data generating process, choosing such an upper bound is more complicated. By (1.2), $\widehat{\beta}_t$ must be in the ℓ_1-ball with radius t. This constraint is only binding Osborne, Presnell and Turlach (2000) if $t < \min_{\eta \in \mathcal{K}} ||\widehat{\beta}_\infty + \eta||_1 =: t_0$, where $\widehat{\beta}_\infty = \widehat{\beta}(\mathbb{R}^p)$ is a least-squares solution and $\mathcal{K} := \{a : \mathbb{X}a = 0\}$. Observe that $\mathcal{K} = \{0\}$ if $n \geq p$ and otherwise $\mathcal{K}$ has dimension $p - n$, which implies $\widehat{\beta}_\infty$ is not unique. Both of these statements assume that the columns of $\mathbb{X}$ contain a linearly independent set of size $\min\{n, p\}$. See Tibshirani (2013) for the more general situation. In either case, if $t \geq t_0$, then $\widehat{\beta}_t$ is equal to a least-squares solution. Choosing the upper bound to be $t_{\max} := ||\widehat{\beta}_\infty||_1$, where $\widehat{\beta}_\infty = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$ is the least-squares solution when $(\cdot)^\dagger$ is given by the Moore-Penrose inverse, suffers from numerical and practical implementation issues if $\operatorname{rank}(\mathbb{X}) < p$. As well it grows much too fast, at least order $\sqrt{n}$, therefore potentially including solutions which have low bias but very high variance.

We take an upper bound based on a rudimentary variance estimator $t_{\max} :=$

$||Y||_2^2/a_n$, where (a_n) is an increasing sequence of constants, and hence specify the optimization interval $T = [0, t_{\max}]$. We defer fixing a particular sequence (a_n)

Remark 1. We emphasize here that using a cross-validated tuning parameter involves more than simply allowing the tuning parameter to be selected in a data-dependent manner. In order to meaningfully analyze tuning parameter selection, we allow the search set T and the tuning parameter to be chosen based on the data.

Remark 2. The computational implementation of CV for an interval Λ (or T) deserves some discussion. Two widely used algorithms for lasso are `glmnet` Friedman, Hastie and Tibshirani (2010), which uses coordinate descent, and `lars` Efron et al. (2004), which leverages the piece-wise linearity of the lasso solution as λ varies (the lasso path). The package `glmnet` is much faster than `lars`; `glmnet` examines a grid of values, $\lambda_j \in \Lambda$, $j = 1, \ldots, J$ say, and approximates the solution at each λ_j with increasing accuracy depending on the number of iterations; `lars` evaluates the entire solution path exactly, such that it is theoretically possible to choose any $\lambda \in \Lambda$ via numerical optimization. Optimizing (2.5) with standard solvers can be difficult due to a possible lack of convexity. In both cases, the extremes of the interval Λ affect the quality of the solution.

3. Main Results

In this section, we present results for both forms of the lasso estimator, (1.1) and (1.2), under more and less restrictive assumptions, respectively. To define the types of random variables $\mathcal{Z}$ we allow, we appeal to the notion of an Orlicz norm.

Definition 1. *For any random variable ξ and function ψ that is nondecreasing, convex, and $\psi(0) = 0$, the ψ-Orlicz norm is*

$$||\xi||_\psi := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\xi|}{c}\right) \leq 1\right\}.$$

For any integer $r \geq 1$, we are interested in the L^r-norm $||\xi||_r := (\mathbb{E}|\xi|^r)^{1/r}$ and the norm given by choosing $\psi(x) = \psi_r(x) := \exp(x^r) - 1$, $||\xi||_{\psi_r}$. Note that if $||\xi||_{\psi_r} < \infty$, then for sufficiently large x, there are constants C_1, c_2 such that $P(|\xi| > x) \leq C_1 \exp(-c_2 x^r)$.

In the particular case of the ψ_2-Orlicz norm, if $||\xi||_{\psi_2} < \kappa$ it holds that $\mathbb{P}(|\xi| > x) \leq 2\exp(-x^2/\kappa^2)$ and $\mathbb{E}[|\xi|^k] \leq 2\kappa^k\Gamma(k/2+1)$, where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the Gamma function. Additionally, $(\mathbb{E}|\xi|^r)^{1/r} = ||\xi||_r \leq r!||\xi||_{\psi_1}$ and $||\xi||_{\psi_1} \leq$

$(\log 2)^{1/r-1}||\xi||_{\psi_r}$.

Definition 2. *Let* $1 \leq q < \infty$ *and* C_q *be a constant independent of* n. *Then*

$$\mathcal{F}_q := \left\{ (\mu_n) : \mu_n \textit{ is a measure on } \mathbb{R}^{p_n} \textit{ and } \max_{1 \leq j,k \leq p} ||\xi_j \xi_k - \mathbb{E}_{\mu_n} \xi_j \xi_k||_{\psi_q} \leq C_q \right\};$$

while the set $\mathcal{F}_\infty$ *contains the measures* μ_n *such that* $|\xi_j| \leq C_q$ *almost surely* μ_n *for each* $j = 1, \ldots, p$.

Remark 3. To make subsequent expressions as a function of q make sense, interpret for any $0 < c < \infty$, $c/\infty = 0$ and $\infty/\infty = 1$.

While μ_n is a measure on $\mathbb{R}^p$, indexing with n is more natural than indexing with p given that our results include p_n increasing with n. Definition 2 specifies a common moment condition (Greenshtein and Ritov (2004); Nardi and Rinaldo (2008); Bartlett, Mendelson and Neeman (2012)) for showing risk consistency of lasso-type methods in high dimensional settings.

We make the following condition about the size of the validation sets for CV.

Condition 1. *The sequence of validation sets* $\{V_n\}_{n=1}^{\infty}$ *is such that, as* $n \to \infty$, $\forall v \in V_n$, $|v| \asymp c_n$ *for some sequence* c_n.

Standard CV methods satisfy this. For example, with K-fold cross-validation, we can take $c_n = \lfloor n/K \rfloor$. For n design random variables $X_1, \ldots, X_n$ and oracle prediction function f^*, let $\mathbf{f}_n^* := (f^*(X_1), \ldots, f^*(X_n))^\top$.

3.1. Persistence when f^* is linear

If we are willing to impose strong conditions on μ_n, as in Bunea, Tsybakov and Wegkamp (2007) and Meinshausen (2007), then we can obtain cross-validated lasso estimators which achieve nearly oracle rates.

If $\mathbb{E}[\mathcal{Y} \mid \mathcal{X}] = f^*(\mathcal{X}) = \mathcal{X}^\top \beta^*$, then we can write the risk of an estimator $\widehat{\beta}_{\widehat{\lambda}}$ as

$$R\left(\widehat{\beta}_{\widehat{\lambda}}\right) = \underbrace{R\left(\widehat{\beta}_{\widehat{\lambda}}\right) - R(\beta^*)}_{\text{excess risk}} + \underbrace{R(\beta^*)}_{\text{noise}} = \underbrace{R\left(\widehat{\beta}_{\widehat{\lambda}}\right) - \sigma^2}_{\text{excess risk}} + \underbrace{\sigma^2}_{\text{noise}},$$

where $R(\beta^*) = \mathbb{E}[(\mathcal{Y} - \mathcal{X}^\top \beta^*)^2] = \sigma^2$. We write the *excess risk* as $\mathcal{E}(\widehat{\lambda}) := R(\widehat{\beta}_{\widehat{\lambda}}) - \sigma^2$ and prove a convergence rate for $\mathcal{E}(\widehat{\lambda})$. In this case, targeting the excess risk is the same as estimating the conditional expectation of $\mathcal{Y}$, but if $f^*(\mathcal{X})$ is not linear (as in Section 3.2), the excess risk remains a meaningful way of assessing prediction behavior. We require some conditions.

M1: There exists a constant $C_q < \infty$ independent of n such that $|\mathcal{X}_j| < C_q$ almost surely for all $j = 1, \ldots, p$.

M2: $\mathbb{E}[\mathcal{X}] = 0$ and $\mathbb{E}[\mathcal{X}\mathcal{X}^\top] = \Sigma$.
M3: $\epsilon \sim N(0, \sigma^2)$.
M4: $\exists 0 < \nu < 1$ such that Σ and Σ^{-1} are diagonally dominant at level ν: $|\sigma_{jj}| \geq \nu + \sum_{j \neq i} |\sigma_{ij}|$ for all $1 \leq j \leq p$.
M5: $||\beta^*||_0 = s^*$, independent of n.
M6: $\lambda_{\min} \propto (\log n \log p/n)^{1/2}$.
M7: $\log p = o(n/\log n)$.

Theorem 1. *Under* **M1** - **M7**, $\mathcal{E}(\widehat{\lambda}) = O_p\left((s^* \log n \log p)/n\right)$.

Here, condition **M4** implies that $\Sigma - (1-\nu)\text{diag}(\Sigma)$ is positive semi-definite. As s^* is fixed, **M6** and **M7** ensure that $\lambda_{\min} \to 0$ so that Λ eventually allows models with s^* covariates. Thus, the procedure is asymptotically consistent for model selection Meinshausen (2007). Here, $\mathcal{E}(\widehat{\lambda})$ is random due to the term $R(\widehat{\beta}_{\widehat{\lambda}})$, so we emphasize that $R(\widehat{\beta}_{\widehat{\lambda}})$ is a function of the data: the conditional expectation of a new test random variable $\mathcal{Z}$ given the observed data used to choose both $\widehat{\lambda}$ and $\widehat{\beta}_\lambda$ as in (2.2).

The conditions of Theorem 1 are typical of those used to prove persistence of the lasso estimator with oracle tuning parameter (for the case of fixed design, see Negahban et al. (2012)). For instance, Bunea, Tsybakov and Wegkamp (2007) prove an oracle rate for the lasso of $O(s^* \log p/n)$ with $\lambda_{\min} \propto \sigma\sqrt{\log p/n}$. Under similar conditions, our result with cross-validated tuning parameter requires a larger $\lambda_{\min}$ (resulting in smaller models) and a slower convergence rate to the oracle by a factor $\log n$. A reasonable choice of Λ suggested by Theorem 1 is $\Lambda = [\lambda_{\min},\ \infty) = [(\log p \log n/n)^{1/2},\ \infty)$.

Proof of Theorem 1. We have that, for all $g > 0$,

$$\mathbb{P}\left(R(\widehat{\beta}_{\widehat{\lambda}}) - \sigma^2 > g\frac{s^* \log n \log p}{n}\right) \leq \mathbb{P}\left(\inf_{\lambda \in \Lambda}\left(R(\widehat{\beta}_\lambda) - \sigma^2\right) > g\frac{s^* \log n \log p}{2n}\right)$$
$$+ 2\mathbb{P}\left(\sup_{\lambda \in \Lambda}\left|R(\widehat{\beta}_\lambda) - \sigma^2 - \widehat{R}(\widehat{\beta}_\lambda)\right| > g\frac{s^* \log n \log p}{2n}\right),$$

by the proof of Theorem 7 in Meinshausen (2007), write $\widehat{R}(\widehat{\beta}_\lambda)$ defined in (2.3). The second term on the right hand side goes to 0 by that result. Now

$$\mathbb{P}\left(\inf_{\lambda \in \Lambda}\left(R(\widehat{\beta}_\lambda) - \sigma^2\right) > g\frac{s^* \log n \log p}{n}\right) \leq \mathbb{P}\left(R(\widehat{\beta}_{\lambda_{\min}}) - \sigma^2 > g\frac{s^* \log n \log p}{n}\right).$$

By Corollary 1 in Bunea, Tsybakov and Wegkamp (2007), for any λ we have

$$\mathbb{P}\left(R(\widehat{\beta}_\lambda) - \sigma^2 > g\frac{s^* \lambda^2}{1-\nu}\right) \leq 10p^2 \exp\left(-c_1 n \lambda^2\right) = 10 \exp\left(2\log p - c_1 n \lambda^2\right).$$

Setting $\lambda_{\min}$ proportional to $(\log n \log p/n)^{1/2}$ is enough for the upper bound to go to zero as $n \to \infty$, yielding the result.

3.2. Persistence when f^* is not linear

In order to derive results under more general conditions, we move to the linear oracle estimation framework. For any t, the oracle estimator with respect to t is

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

Suppose $\widehat{t}$ is an estimator, such as by cross-validation. Then we can decompose the risk of an estimator $\widehat{\beta}_{\widehat{t}}$ as

$$R\left(\widehat{\beta}_t\right) = \underbrace{R\left(\widehat{\beta}_{\widehat{t}}\right) - R\left(\beta_t\right)}_{\text{excess risk}} \quad + \quad \underbrace{R\left(\beta_t\right) - R_*}_{\text{approximation error}} \quad + \quad \underbrace{R_*}_{\text{noise}},$$

where R_* is the risk of the mean function f^*. Because the data generating process is not necessarily linear, we study the performance of an estimator $\widehat{\beta}_{\widehat{t}}$ via the *excess risk of* $\widehat{\beta}_{\widehat{t}}$ *relative to* β_t,

$$\mathcal{E}(\widehat{t}, t) := R\left(\widehat{\beta}_{\widehat{t}}\right) - R\left(\beta_t\right). \tag{3.1}$$

Here, $\mathcal{E}(\widehat{t}, t)$ depends on the cross-validated tuning parameter $\widehat{t}$ as well as the oracle set through t. Focusing on (3.1) allows for meaningful theory when the approximation error does not necessarily converge to zero as n grows. This is important as we do not assume that the conditional expectation of $\mathcal{Y}$ given $\mathcal{X}$ is linear. As $t \to \infty$, the approximation error decreases and hence we desire to take $t = t_n$ for some increasing sequence (t_n). Greenshtein and Ritov (2004) show that if $t_n = o((n/\log p)^{1/4})$, then $\mathcal{E}(t_n, t_n)$ converges in probability to zero. Bartlett, Mendelson and Neeman (2012) increase the size of this search set allowing $t_n = o(n^{1/2}/(\log^{3/2} n \log^{3/2}(np)))$ while still having $\mathcal{E}(t_n, t_n) \xrightarrow{P} 0$.

Theorem 2. *Let* $(\mu_n) \in \mathcal{F}_q$ *and suppose that Condition* 1 *holds. Then for any sequences* (a_n), (t_n) *which satisfy* $a_n t_n = o(n)$,

$$\mathbb{P}_{\mu_n}\left(\mathcal{E}(\widehat{t}, t_n) > \delta\right) \leq \delta^{-1}\left(\Omega_{n,1} + \Omega_{n,2}\right) + 2\mathbb{P}(D_n^c) + \mathbb{P}(E_n^c), \tag{3.2}$$

where,

$$\Omega_{n,1} := \left[1 + \frac{2nC_q'}{a_n}\right]^2 \sqrt{(\log p)^{1+2/q}} \left(n^{-1/2} + c_n^{-1/2} + (n - c_n)^{-1/2}\right),$$

$$\Omega_{n,2} := (1 + t_n)^2 \sqrt{\frac{(\log p)^{1+2/q}}{n}},$$

$$
\begin{aligned}
D_n &:= \left\{ t_{\max} \leq \frac{2nC'_q}{a_n} \right\}, \\
E_n &:= \{t_{\max} \geq t_n\}, \\
C'_q &= C_q(\log 2)^{1/q-1}.
\end{aligned}
$$

Remark 4. The sets D_n, E_n account for cases wherein $t_{\max} = ||Y||_2 / a_n$ results in suboptimal sets T. If (a_n) is such that $t_{\max}$ grows too quickly with non-negligible probability, then cross-validation may result in low-bias but high variance solutions. On the other hand, if $t_{\max}$ is too small, then we rule out possible estimators with lower risk. Here D_n calibrates a high-probability upper bound on $t_{\max}$ based on (μ_n) and the choice of (a_n) while E_n ensures that $t_{\max}$ will be large enough to include low risk estimators.

Remark 5. Usually in the oracle estimation framework, $\widehat{t} = t_n$ and so the excess risk is necessarily nonnegative because the oracle predictor, β_{t_n}, is selected as the risk minimizer over $\mathcal{B}_{t_n}$. In that case, (3.2) corresponds to convergence in probability. As we are examining the case where the optimization set is estimated, $\mathcal{E}(\widehat{t}, t_n)$ may be negative. However, we are only interested in the case where the estimator is worse than the oracle.

Let $b_n = \min\{n - c_n, c_n\}$, then $(n^{-1/2} + c_n^{-1/2} + (n - c_n)^{-1/2}) \leq 3b_n^{-1/2}$.

Corollary 1. *Under the conditions of Theorem* 2*, if* $a_n = n(\log p)^{1/4+1/(2q)} m_n / b_n^{1/4}$ *and* $t_n = o(b_n^{1/4}/m_n(\log p)^{1/4+1/(2q)})$*, where* m_n *is any sequence which tends toward infinity and* $m_n = o(b_n^{1/4})$*, for* n *large enough and* $p = o(\exp\{b_n^{q/(q+2)}\})$*,*

$$
\mathbb{P}_{\mu_n}\left(\mathcal{E}(\widehat{t}, t_n) > \delta\right) \leq \frac{1}{m_n^2 \delta}\left(1 + \sqrt{\frac{b_n}{n}}\right) + 2\exp(-\frac{n}{8e^2}).
$$

In particular, $\mathbb{P}_{\mu_n}\left(\mathcal{E}(\widehat{t}, t_n) > \delta\right) \to 0$.

Remark 6. The rate at which δ can be taken to zero quantifies the decay of the size of the 'bad' set where $\mathcal{E}(\widehat{t}, t_n)$ is large. For the corollary, both $m_n = o(b_n^{1/4})$ and $\delta^{-1} = o(m_n^2)$. Therefore, it is necessary for $\delta^{-1} = o(b_n^{1/2})$ and hence, δ must go to zero at a slower rate than $b_n^{-1/2}$.

Remark 7. As q increases, which corresponds to $(\mu_n) \in \mathcal{F}_q$ putting less mass on the tails of the centered interactions of the components of $\mathcal{Z}$, the faster the oracle set given by $\mathcal{B}_{t_n}$ can grow. When $q = \infty$, the random variables have no tails and we get the fastest rate of growth for t_n, $(b_n^{1/4}/(m_n(\log p)^{1/4}))$.

The parameter b_n controls the minimum size of the validation versus training sets that comprise cross-validation. It must be true that b_n is strictly less than

n. To get the best guarantee, b_n should increase as fast as possible. Hence, our results advocate a cross-validation scheme where the validation and training sets are asymptotically balanced, $c_n \asymp n - c_n$. This should be compared with the results in Shao (1993), which state that for model selection consistency, one should have $c_n/n \to 1$. However, Shao (1993) presents results for model selection while we focus on prediction error. Similarly, Dudoit and van der Laan (2005) provide oracle inequalities for cross-validation and also advocate for the validation set to grow as fast as possible. For K-fold cross-validation, $c_n = \lfloor n/K \rfloor$ so that $b_n = O(n)$.

It is instructive to compare this choice of $t_{\max}$ with $||Y||_2^2/n$, a standard estimate of the noise variance in high dimensions (e.g. (van de Geer and Bühlmann, 2011, p.104)). If $a_n = n$, then $||Y||_2^2/a_n$ is an overestimate of the variance due to the presence of the regression function f^*. Our results state that we must choose a_n to increase slower than n, thereby increasing this overestimation and enlarging the potential search set T. While a_n depends on several parameters, b_n, n, and p are known to the analyst. Also, the choice of q depends on how much approximation error the analyst is willing to make. The m_n term is required to slow the growth of t_n just slightly. While this shrinks the size of the set $\mathcal{B}_{t_n}$ relative to that used by Greenshtein and Ritov (2004), potentially eliminating some better solutions, effectively m_n quantifies their requirement $t_n = o((n/\log p)^{1/4})$, making explicit the need for t_n to grow more slowly than $(n/\log p)^{1/4}$. As such, if we set $b_n \asymp n$ and set $q = \infty$, we require the rate shown by Greenshtein and Ritov (2004), where Corollary 1 implies that both $\mathbb{P}_{\mu_n}\left(\mathcal{E}(\widehat{t}, t_n) > \delta\right) \to 0$ and $t_n = o((n/\log p)^{1/4})$.

Proofs of Theorem 2 and Corollary 1 are in the supplementary material.

Our results generalize to other M-estimators which use an ℓ_1-constraint. In particular, relative to the set of coefficients $\beta \in \mathcal{B}_{t_n}$ with $t_n = o(b_n^{1/4}/(m_n(\log p)^{1/4+1/(2q)}))$, an empirical estimator with cross-validated tuning parameter has a prediction risk that converges to the prediction risk of the oracle.

Corollary 2. *For the group lasso* Yuan and Lin (2006)

$$\widehat{\beta}_t = \operatorname{argmin}\{n^{-1}\,||Y - \mathbb{X}\beta||_2^2 : \sum_{g \in G} \sqrt{|g|}\,||\beta_g||_2 \le t\},$$

and the square-root lasso Belloni, Chernozhukov and Wang (2014)

$$\widehat{\beta}_t = \operatorname{argmin}\{n^{-1}\,||Y - \mathbb{X}\beta||_2 : ||\beta||_1 \le t\},$$

if t_n and a_n are as in Corollary 1*, then, for n large enough,* $\log(p) = o(b_n^{q/(q+2)})$*, and* $\max_g \sqrt{|g|} = O(1)$*, we have*

$$\mathbb{P}_{\mu_n}\left(\mathcal{E}(\widehat{t}, t_n) > \delta\right) \leq \frac{1}{m_n^2 \delta}\left(1 + \sqrt{\frac{b_n}{n}}\right) + 2e^{-n/8e^2}.$$

4. Simulations

We simulated the predictive and model selection performance of K-fold cross-validation for a range of K.

We considered three criteria: prediction risk, sensitivity, and specificity. For prediction risk, we approximated $R(\widehat{\beta}_{\widehat{\lambda}})$ by the empirical risk of 500 test observations. For sensitivity and specificity, the active set of a coefficient vector β was $\mathcal{S}(\beta) := \{j : |\beta_j| > 0\}$, with $\mathcal{S}^* := \mathcal{S}(\beta^*)$ and $\widehat{\mathcal{S}} := \mathcal{S}(\widehat{\beta}_{\widehat{\lambda}})$ as the active sets of β^* and $\widehat{\beta}_{\widehat{\lambda}}$, respectively. Thus, sensitivity $= |\mathcal{S}^* \cap \widehat{\mathcal{S}}|/|\mathcal{S}^*|$ and specificity $= |(\mathcal{S}^*)^c \cap \widehat{\mathcal{S}}^c|/|(\mathcal{S}^*)^c|$, where $|\cdot|$ counts the number of elements in a set and $\mathcal{A}^c$ indicates the complement of a set $\mathcal{A}$.

4.1. Simulation details

Conditions: We considered a wide range of possible conditions by varying the correlation in the design, ρ; the number of parameters, p; the sparsity, α; and the signal-to-noise ratio, SNR. In all cases, we let $n = 100$, $p = 75, 350, 1,000$, and set the measurement error variance $\sigma^2 = 1$. We ran each simulation condition combination 100 times, assuming that there existed a β^* such that the regression function $f^*(X) = X^\top \beta^*$ in order to make model selection meaningful.

The design matrices were produced in two steps. First, $X_{ij} \overset{i.i.d.}{\sim} N(0, 1)$ for $1 \leq i \leq n$ and $1 \leq j \leq p$, formed the matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$. Second, correlations were introduced by defining a matrix D with all off-diagonal elements equal to ρ and diagonal elements equal to one. Then, we took $\mathbb{X} \leftarrow \mathbb{X}D^{1/2}$. For these simulations, we considered correlations $\rho = 0.2, 0.5, 0.95$.

For sparsity, we took $s^* = \lceil n^\alpha \rceil$ and generateg the s^* non-zero elements of β^* from a Laplace distribution with parameter 1. We let $\alpha = 0.1, 0.33, 0.5$. To compensate for the random amount of signal in each observation, we varied the signal-to-noise ratio, SNR $= \beta^\top D\beta$. We considered SNR $= 1$ and 10. As the SNR increases the observations go from a high-noise and low-signal regime to a low-noise and high-signal one. We considered $\epsilon \sim N(0, 1)$ and $\epsilon \sim 3^{-1/2}t(3)$, where $t(3)$ is the t distribution with 3 degrees of freedom and the $3^{-1/2}$ term makes the variance equal to 1. Finally, we took $K = \{3,\ 10,\ 25,\ 50,\ 75,\ 100\}$, the last case being leave-one-out CV.

4.2. Simulation results

Of the simulations, we have only included the most informative plots. For

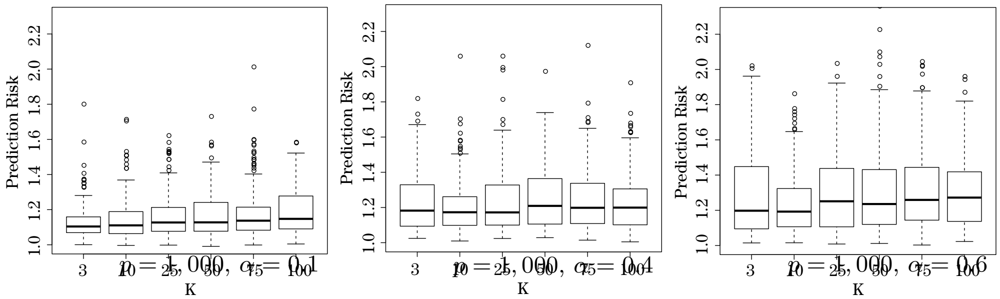

Figure 1. **Prediction risk:** The other parameters are set at: SNR = 1, $\rho = 0.9$.

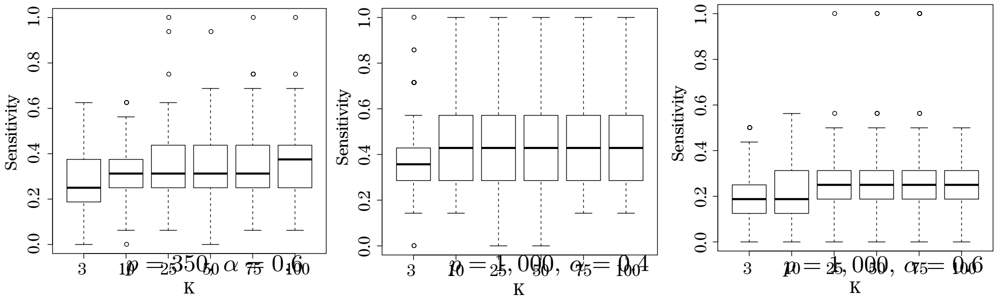

Figure 2. **Sensitivity:** The other parameters are set at: SNR = 1, $\rho = 0.2$.

prediction risk, all considered K's resulted in remarkably similar prediction risks. In Figure 1, for $p = 1,000$, SNR = 1, and $\rho = 0.9$, we see that taking $K = 3$ or $K = 10$ results in slightly smaller prediction risks. This is comforting as both of these values of K are used frequently by default and they are the least computationally demanding.

For model selection, there is a more nuanced story. For sensitivity, which describes how often we would correctly identify a coefficient as nonzero, larger values of K tended to work better. For instance, in Figure 2, we see that $K = 3$ is often decidedly worse than larger K, followed by $K = 10$. As is often the case, this conclusion presents a trade-off with the results for specificity (Figure 3): smaller values of K tended to work better. In general, $\widehat{\beta}(\widehat{\lambda})$ tended to have more nonzero entries as K increased holding all else constant. As the correlation parameter (ρ) or the signal to noise (SNR) increased, all values of K had approximately the same performance.

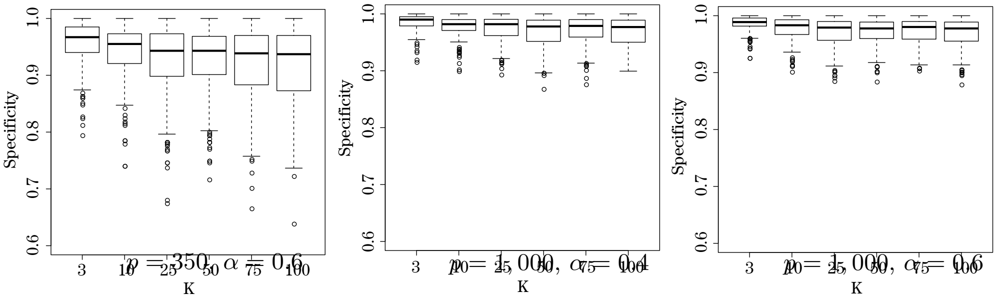

Figure 3. **Specificity:** The other parameters are set at: SNR = 1, $\rho = 0.2$.

5. Discussion

Our work leaves some interesting open questions. Our most general results do not apply for leave-one-out cross-validation as $c_n = 1$ for all n, and hence the upper-bounds become trivial. Leave-one-out cross-validation is more computationally demanding than K-fold cross-validation, but is still used in practice. Our results do not give any prescription for choosing K other than that it should be $o(n)$. Our simulation study indicates that all K ranging from 3 to n have approximately the same predictive ability. For model selection, larger K tends to produce more nonzero coefficients and hence has better sensitivity but poorer specificity.

As there are many other methods for choosing the tuning parameter in the lasso problem, a direct comparison of the behavior of the lasso estimator with tuning parameter chosen via cross-validation versus a degrees-of-freedom-based method is of substantial interest. Our results depend strongly on the upper bound for T or the lower bound for Λ, but, in most cases, we never need to use tuning parameters this extreme. It makes sense to attempt to find more subtle theory to provide greater intuition for the behavior of lasso under cross-validation.

Supplementary Materials

Supplementary materials include proofs of theorems and lemmata.

Acknowledgment

The authors gratefully acknowledge support from the National Science Foundation (DH, DMS–1407543; DJM, DMS–1407439) and the Institute of New Economic Thinking (DH and DJM, INO14-00020). Additionally, the authors thank Cosma Shalizi and Ryan Tibshirani for reading preliminary versions of this manuscript and, Larry Wasserman for inspiration and encouragement.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**, 40–79.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research* **10**, 245–279.

Bartlett, P. L., Mendelson, S. and Neeman, J. (2012). ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probability Theory and Related Fields* **154**, 193–224.

Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics* **42**, 757–788.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer New York.

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

Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics* **37**, 2145–2177.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* **22**, 555–574.

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

Donoho, D. L., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory* **52**, 6–18.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validation risk estimation in estimator selection and performance assessment. *Statistical Methodology* **2**, 131–154.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society B* **75**, 531–552.

Flynn, C. J., Hurvich, C. M. and Simonoff, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association* **108**, 1031–1043.

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

Giraud, C., Huet, S. and Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statistical Science* **27**, 500–518.

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. (2013). The lasso, persistence, and cross-validation. in *Proceedings of The 30th International Conference on Machine Learning* pp. 1031–1039.

Homrighausen, D. and McDonald, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning* **97**, 65–78.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis* **100**, 1338–1352.

Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research* **13**, 1037–1057.

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

Lecué, G. and Mitchell, C. (2012). Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics* **6**, 1803–18374.

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

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

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

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

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* **2**, 605–633.

Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**, 538–337.

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

Saumard, A. (2011). The slope heuristics in heteroscedastic regression. arXiv:1104.1050.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64**, 29–35.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.

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

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7**, 1456–1490.

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

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

van de Geer, S. A. and Bühlmann, P. (2011). *Statistics for High-Dimensional Data*, Springer Verlag.

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

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

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 671–683.

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

Zhang, Y. and Shen, X. (2010). Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **3**, 350–358.

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

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

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

Department of Statistics, Colorado State University, Fort Collins, Colorado, 80523 USA

E-mail: darrenho@stat.colostate.edu

Department of Statistics, Indiana University, Bloomington, IN 47405, USA

E-mail: dajmcdon@indiana.edu

(Received October 2015; accepted May 2016)