
Bolasso: Model Consistent Lasso Estimation through the Bootstrap

Francis R. Bach

FRANCIS.BACH@MINES.ORG

INRIA - WILLOW Project-Team

Laboratoire d'Informatique de l'École Normale Supérieure (CNRS/ENS/INRIA UMR 8548)

45 rue d'Ulm, 75230 Paris, France

Abstract

We consider the least-square linear regression problem with regularization by the ℓ_1 -norm, a problem usually referred to as the Lasso. In this paper, we present a detailed asymptotic analysis of model consistency of the Lasso. For various decays of the regularization parameter, we compute asymptotic equivalents of the probability of correct model selection (i.e., variable selection). For a specific rate decay, we show that the Lasso selects all the variables that should enter the model with probability tending to one exponentially fast, while it selects all other variables with strictly positive probability. We show that this property implies that if we run the Lasso for several bootstrapped replications of a given sample, then intersecting the supports of the Lasso bootstrap estimates leads to consistent model selection. This novel variable selection algorithm, referred to as the Bolasso, is compared favorably to other linear regression methods on synthetic data and datasets from the UCI machine learning repository.

1. Introduction

Regularization by the ℓ_1 -norm has attracted a lot of interest in recent years in machine learning, statistics and signal processing. In the context of least-square linear regression, the problem is usually referred to as the *Lasso* (Tibshirani, 1994). Much of the early effort has been dedicated to algorithms to solve the optimization problem efficiently. In particular, the *Lars* algorithm of Efron et al. (2004) allows to find the entire regularization path (i.e., the set of solutions for all values of the regularization parameters) at the cost of

a single matrix inversion.

Moreover, a well-known justification of the regularization by the ℓ_1 -norm is that it leads to *sparse* solutions, i.e., loading vectors with many zeros, and thus performs model selection. Recent works (Zhao & Yu, 2006; Yuan & Lin, 2007; Zou, 2006; Wainwright, 2006) have looked precisely at the model consistency of the Lasso, i.e., if we know that the data were generated from a sparse loading vector, does the Lasso actually recover it when the number of observed data points grows? In the case of a fixed number of covariates, the Lasso does recover the sparsity pattern if and only if a certain simple condition on the generating covariance matrices is verified (Yuan & Lin, 2007). In particular, in low correlation settings, the Lasso is indeed consistent. However, in presence of strong correlations, the Lasso cannot be consistent, shedding light on potential problems of such procedures for variable selection. Adaptive versions where data-dependent weights are added to the ℓ_1 -norm then allow to keep the consistency in all situations (Zou, 2006).

In this paper, we first derive a detailed asymptotic analysis of sparsity pattern selection of the Lasso estimation procedure, that extends previous analysis (Zhao & Yu, 2006; Yuan & Lin, 2007; Zou, 2006), by focusing on a specific decay of the regularization parameter. We show that when the decay is proportional to $n^{-1/2}$, where n is the number of observations, then the Lasso will select all the variables that should enter the model (the *relevant* variables) with probability tending to one exponentially fast with n , while it selects all other variables (the *irrelevant* variables) with strictly positive probability. If several datasets generated from the same distribution were available, then the latter property would suggest to consider the intersection of the supports of the Lasso estimates for each dataset: all relevant variables would always be selected for all datasets, while irrelevant variables would enter the models randomly, and intersecting the supports from sufficiently many different datasets would simply eliminate them. However, in practice, only one

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

dataset is given; but resampling methods such as the *bootstrap* are exactly dedicated to mimic the availability of several datasets by resampling from the same unique dataset (Efron & Tibshirani, 1998). In this paper, we show that when using the bootstrap and intersecting the supports, we actually get a consistent model estimate, without the consistency condition required by the regular Lasso. We refer to this new procedure as the *Bolasso* (**boot**strap-enhanced **least absolute shrinkage operator**). Finally, our Bolasso framework could be seen as a voting scheme applied to the supports of the bootstrap Lasso estimates; however, our procedure may rather be considered as a consensus combination scheme, as we keep the (largest) subset of variables on which *all* regressors agree in terms of variable selection, which is in our case provably consistent and also allows to get rid of a potential additional hyperparameter.

The paper is organized as follows: in Section 2, we present the asymptotic analysis of model selection for the Lasso; in Section 3, we describe the Bolasso algorithm as well as its proof of model consistency, while in Section 4, we illustrate our results on synthetic data, where the true sparse generating model is known, and data from the UCI machine learning repository. Sketches of proofs can be found in Appendix A.

Notations For a vector $v \in \mathbb{R}^p$, we let denote $\|v\|_2 = (v^\top v)^{1/2}$ the ℓ_2 -norm, $\|v\|_\infty = \max_{i \in \{1, \dots, p\}} |v_i|$ the ℓ_∞ -norm and $\|v\|_1 = \sum_{i=1}^p |v_i|$ the ℓ_1 -norm. For $a \in \mathbb{R}$, $\text{sign}(a)$ denotes the sign of a , defined as $\text{sign}(a) = 1$ if $a > 0$, -1 if $a < 0$, and 0 if $a = 0$. For a vector $v \in \mathbb{R}^p$, $\text{sign}(v) \in \mathbb{R}^p$ denotes the the vector of signs of elements of v .

Moreover, given a vector $v \in \mathbb{R}^p$ and a subset I of $\{1, \dots, p\}$, v_I denotes the vector in $\mathbb{R}^{\text{Card}(I)}$ of elements of v indexed by I . Similarly, for a matrix $A \in \mathbb{R}^{p \times p}$, $A_{I,J}$ denotes the submatrix of A composed of elements of A whose rows are in I and columns are in J .

2. Asymptotic Analysis of Model Selection for the Lasso

In this section, we describe existing and new asymptotic results regarding the model selection capabilities of the Lasso.

2.1. Assumptions

We consider the problem of predicting a response $Y \in \mathbb{R}$ from covariates $X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$. The only assumptions that we make on the joint distribution P_{XY} of (X, Y) are the following:

- (A1) The cumulant generating functions $\mathbb{E} \exp(s\|X\|_2^2)$ and $\mathbb{E} \exp(sY^2)$ are finite for some $s > 0$.
- (A2) The joint matrix of second order moments $\mathbf{Q} = \mathbb{E}XX^\top \in \mathbb{R}^{p \times p}$ is invertible.
- (A3) $\mathbb{E}(Y|X) = X^\top \mathbf{w}$ and $\text{var}(Y|X) = \sigma^2$ a.s. for some $\mathbf{w} \in \mathbb{R}^p$ and $\sigma \in \mathbb{R}_+^*$.

We let denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the sparsity pattern of \mathbf{w} , $\mathbf{s} = \text{sign}(\mathbf{w})$ the sign pattern of \mathbf{w} , and $\varepsilon = Y - X^\top \mathbf{w}$ the additive noise.¹ Note that our assumption regarding cumulant generating functions is satisfied when X and ε have compact support, and also, when the densities of X and ε have light tails.

We consider *independent and identically distributed* (i.i.d.) data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, sampled from P_{XY} ; the data are given in the form of matrices $\bar{Y} \in \mathbb{R}^n$ and $\bar{X} \in \mathbb{R}^{n \times p}$.

Note that the i.i.d. assumption, together with (A1-3), are the simplest assumptions for studying the asymptotic behavior of the Lasso; and it is of course of interest to allow more general assumptions, in particular growing number of variables p , more general random variables, etc. (see, e.g., Meinshausen and Yu (2006)), which are outside the scope of this paper.

2.2. Lasso Estimation

We consider the square loss function $\frac{1}{2n} \sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{2n} \|\bar{Y} - \bar{X}w\|_2^2$ and the regularization by the ℓ_1 -norm defined as $\|w\|_1 = \sum_{i=1}^p |w_i|$. That is, we look at the following Lasso optimization problem (Tibshirani, 1994):

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\bar{Y} - \bar{X}w\|_2^2 + \mu_n \|w\|_1, \quad (1)$$

where $\mu_n \geq 0$ is the regularization parameter. We denote \hat{w} any global minimum of Eq. (1)—it may not be unique in general, but will with probability tending to one exponentially fast under assumption (A1).

2.3. Model Consistency - General Results

In this section, we detail the asymptotic behavior of the Lasso estimate \hat{w} , both in terms of the difference in norm with the population value \mathbf{w} (i.e., regular consistency) and of the *sign pattern* $\text{sign}(\hat{w})$, for all asymptotic behaviors of the regularization parameter μ_n . Note that information about the sign pattern includes information about the *support*, i.e., the indices i for which \hat{w}_i is different from zero; moreover, when \hat{w} is consistent, consistency of the sign pattern is in fact equivalent to the consistency of the support.

¹Throughout this paper, we use boldface fonts for population quantities.

We now consider five mutually exclusive possible situations which explain various portions of the regularization path (we assume **(A1-3)**); many of these results appear elsewhere (Yuan & Lin, 2007; Zhao & Yu, 2006; Fu & Knight, 2000; Zou, 2006; Bach, 2007) but some of the finer results presented below are new (see Section 2.4).

1. If μ_n tends to infinity, then $\hat{w} = 0$ with probability tending to one.
2. If μ_n tends to a finite strictly positive constant μ_0 , then \hat{w} converges in probability to the unique global minimum of $\frac{1}{2}(w - \mathbf{w})^\top \mathbf{Q}(w - \mathbf{w}) + \mu_0 \|w\|_1$. Thus, the estimate \hat{w} never converges in probability to \mathbf{w} , while the sign pattern tends to the one of the previous global minimum, which may or may not be the same as the one of \mathbf{w} .²
3. If μ_n tends to zero slower than $n^{-1/2}$, then \hat{w} converges in probability to \mathbf{w} (regular consistency) and the sign pattern converges to the sign pattern of the global minimum of $\frac{1}{2}v^\top \mathbf{Q}v + v_{\mathbf{J}}^\top \text{sign}(\mathbf{w}_{\mathbf{J}}) + \|v_{\mathbf{J}^c}\|_1$. This sign pattern is equal to the population sign vector $\mathbf{s} = \text{sign}(\mathbf{w})$ if and only if the following consistency condition is satisfied:

$$\|\mathbf{Q}_{\mathbf{J}^c \mathbf{J}} \mathbf{Q}_{\mathbf{J} \mathbf{J}}^{-1} \text{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leq 1. \quad (2)$$

Thus, if Eq. (2) is satisfied, the probability of correct sign estimation is tending to one, and to zero otherwise (Yuan & Lin, 2007).

4. If $\mu_n = \mu_0 n^{-1/2}$ for $\mu_0 \in (0, \infty)$, then the sign pattern of \hat{w} agrees on \mathbf{J} with the one of \mathbf{w} with probability tending to one, while for all sign patterns consistent on \mathbf{J} with the one of \mathbf{w} , the probability of obtaining this pattern is tending to a limit in $(0, 1)$ (in particular strictly positive); that is, all patterns consistent on \mathbf{J} are possible with positive probability. See Section 2.4 for more details.
5. If μ_n tends to zero faster than $n^{-1/2}$, then \hat{w} is consistent (i.e., converges in probability to \mathbf{w}) but the support of \hat{w} is equal to $\{1, \dots, p\}$ with probability tending to one (the signs of variables in \mathbf{J}^c may be negative or positive). That is, the ℓ_1 -norm has no sparsifying effect.

Among the five previous regimes, the only ones with consistent estimates (in norm) and a sparsity-inducing effect are μ_n tending to zero and $\mu_n n^{1/2}$ tending to

a limit $\mu_0 \in (0, \infty]$ (i.e., potentially infinite). When $\mu_0 = +\infty$, then we can only hope for model consistent estimates if the consistency condition in Eq. (2) is satisfied. This somewhat disappointing result for the Lasso has led to various improvements on the Lasso to ensure model consistency even when Eq. (2) is not satisfied (Yuan & Lin, 2007; Zou, 2006). Those are based on adaptive weights based on the non regularized least-square estimate. We propose in Section 3 an alternative way which is based on resampling.

In this paper, we now consider the specific case where $\mu_n = \mu_0 n^{-1/2}$ for $\mu_0 \in (0, \infty)$, where we derive new asymptotic results. Indeed, in this situation, we get the correct signs of the relevant variables (those in \mathbf{J}) with probability tending to one, but we also get all possible sign patterns consistent with this, i.e., all other variables (those not in \mathbf{J}) may be non zero with asymptotically strictly positive probability. However, if we were to repeat the Lasso estimation for many datasets obtained from the same distribution, we would obtain for each μ_0 , a set of active variables, all of which include \mathbf{J} with probability tending to one, but potentially containing all other subsets. By intersecting those, we would get exactly \mathbf{J} .

However, this requires multiple copies of the samples, which are not usually available. Instead, we consider bootstrapped samples which exactly mimic the behavior of having multiple copies. See Section 3 for more details.

2.4. Model Consistency with Exact Root- n Regularization Decay

In this section we present detailed new results regarding the pattern consistency for μ_n tending to zero exactly at rate $n^{-1/2}$ (see proofs in Appendix A):

Proposition 1 *Assume **(A1-3)** and $\mu_n = \mu_0 n^{-1/2}$, $\mu_0 > 0$. Then for any sign pattern $s \in \{-1, 0, 1\}^p$ such that $s_{\mathbf{J}} = \text{sign}(\mathbf{w}_{\mathbf{J}})$, $\mathbb{P}(\text{sign}(\hat{w}) = s)$ tends to a limit $\rho(s, \mu_0) \in (0, 1)$, and we have:*

$$\mathbb{P}(\text{sign}(\hat{w}) = s) - \rho(s, \mu_0) = O(n^{-1/2} \log n).$$

Proposition 2 *Assume **(A1-3)** and $\mu_n = \mu_0 n^{-1/2}$, $\mu_0 > 0$. Then, for any pattern $s \in \{-1, 0, 1\}^p$ such that $s_{\mathbf{J}} \neq \text{sign}(\mathbf{w}_{\mathbf{J}})$, there exist a constant $A(\mu_0) > 0$ such that*

$$\log \mathbb{P}(\text{sign}(\hat{w}) = s) \leq -nA(\mu_0) + O(n^{-1/2}).$$

The last two propositions state that we get all relevant variables with probability tending to one *exponentially fast*, while we get exactly all other patterns with probability tending to a limit *strictly* between zero

²Here and in the third regime, we do not take into account the pathological cases where the sign pattern of the limit is unstable, i.e., the limit is exactly at a hinge point of the regularization path.

and one. Note that the results that we give in this paper are valid for *finite* n , i.e., we could derive actual bounds on probability of sign pattern selections with known constants that explicitly depend on \mathbf{w} , \mathbf{Q} and P_{XY} .

3. Bolasso: Bootstrapped Lasso

Given the n i.i.d. observations $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, n$, given by matrices $\bar{X} \in \mathbb{R}^{n \times p}$ and $\bar{Y} \in \mathbb{R}^n$, we consider m *bootstrap* replications of the n data points (Efron & Tibshirani, 1998); that is, for $k = 1, \dots, m$, we consider a *ghost sample* $(x_i^k, y_i^k) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, n$, given by matrices $\bar{X}^k \in \mathbb{R}^{n \times p}$ and $\bar{Y}^k \in \mathbb{R}^n$. The n pairs (x_i^k, y_i^k) , $i = 1, \dots, n$, are sampled uniformly at random *with replacement* from the n original pairs in (\bar{X}, \bar{Y}) . The sampling of the nm pairs of observations is independent. In other words, we defined the distribution of the ghost sample (\bar{X}^*, \bar{Y}^*) by sampling n points with replacement from (\bar{X}, \bar{Y}) , and, given (\bar{X}, \bar{Y}) , the m ghost samples are independently sampled i.i.d. from the distribution of (\bar{X}^*, \bar{Y}^*) .

The asymptotic analysis from Section 2 suggests to estimate the supports $J_k = \{j, \hat{w}_j^k \neq 0\}$ of the Lasso estimates \hat{w}^k for the bootstrap samples, $k = 1, \dots, m$, and to intersect them to define the Bolasso model estimate of the support: $J = \bigcap_{k=1}^m J_k$. Once J is selected, we estimate w by the unregularized least-square fit restricted to variables in J . The detailed algorithm is given in Algorithm 1. The algorithm has only one extra parameter (the number of bootstrap samples m). Following Proposition 3, $\log(m)$ should be chosen growing with n asymptotically slower than n . In simulations, we always use $m = 128$ (except in Figure 3, where we exactly study the influence of m).

Algorithm 1 Bolasso

Input: data $(\bar{X}, \bar{Y}) \in \mathbb{R}^{n \times (p+1)}$
 number of bootstrap replicates m
 regularization parameter μ
for $k = 1$ **to** m **do**
 Generate bootstrap samples $(\bar{X}^k, \bar{Y}^k) \in \mathbb{R}^{n \times (p+1)}$
 Compute Lasso estimate \hat{w}^k from (\bar{X}^k, \bar{Y}^k)
 Compute support $J_k = \{j, \hat{w}_j^k \neq 0\}$
end for
 Compute $J = \bigcap_{k=1}^m J_k$
 Compute \hat{w}_J from (\bar{X}_J, \bar{Y})

Note that in practice, the Bolasso estimate can be computed simultaneously for a large number of regularization parameters because of the efficiency of the Lasso algorithm (which we use in simulations), that allows to find the entire regularization path for the Lasso at

the (empirical) cost of a single matrix inversion (Efron et al., 2004). Thus computational complexity of the Bolasso is $O(m(p^3 + p^2n))$.

The following proposition (proved in Appendix A) shows that the previous algorithm leads to consistent model selection.

Proposition 3 *Assume (A1-3) and $\mu_n = \mu_0 n^{-1/2}$, $\mu_0 > 0$. Then the probability that the Bolasso does not exactly select the correct model, i.e., for all $m > 0$, $\mathbb{P}(J \neq \mathbf{J})$ has the following upper bound:*

$$\mathbb{P}(J \neq \mathbf{J}) \leq mA_1 e^{-A_2 n} + A_3 \frac{\log n}{n^{1/2}} + A_4 \frac{\log m}{m},$$

where A_1, A_2, A_3, A_4 are strictly positive constants.

Therefore, if $\log(m)$ tends to infinity slower than n when n tends to infinity, the Bolasso asymptotically selects with overwhelming probability the correct active variable, and by regular consistency of the restricted least-square estimate, the correct sign pattern as well. Note that the previous bound is true whether the condition in Eq. (2) is satisfied or not, but could be improved on if we suppose that Eq. (2) is satisfied. See Section 4.1 for a detailed comparison with the Lasso on synthetic examples.

4. Simulations

In this section, we illustrate the consistency results obtained in this paper with a few simple simulations on synthetic examples similar to the ones used by Bach (2007) and some medium scale datasets from the UCI machine learning repository (Asuncion & Newman, 2007).

4.1. Synthetic examples

For a given dimension p , we sampled $X \in \mathbb{R}^p$ from a normal distribution with zero mean and covariance matrix generated as follows: (a) sample a $p \times p$ matrix G with independent standard normal distributions, (b) form $\mathbf{Q} = GG^\top$, (c) scale \mathbf{Q} to unit diagonal. We then selected the first $\text{Card}(\mathbf{J}) = r$ variables and sampled non zero loading vectors as follows: (a) sample each loading from independent standard normal distributions, (b) rescale those to unit magnitude, (c) rescale those by a scaling which is uniform at random between $\frac{1}{3}$ and 1 (to ensure $\min_{j \in \mathbf{J}} |\mathbf{w}_j| \geq 1/3$). Finally, we chose a constant noise level σ equal to 0.1 times $(\mathbb{E}(\mathbf{w}^\top X)^2)^{1/2}$, and the additive noise ε is normally distributed with zero mean and variance σ^2 . Note that the joint distribution on (X, Y) thus defined satisfies with probability one (with respect to the sampling of the covariance matrix) assumptions (A1-3).

In Figure 1, we sampled two distributions P_{XY} with

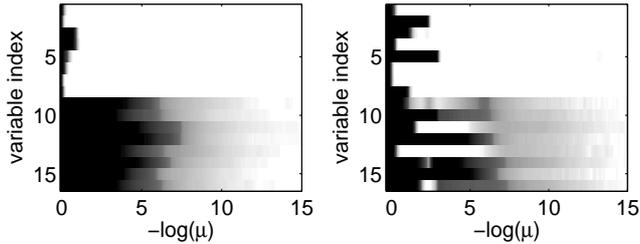


Figure 1. **Lasso**: log-odd ratios of the probabilities of selection of each variable (white = large probabilities, black = small probabilities) vs. regularization parameter. Consistency condition in Eq. (2) satisfied (left) and not satisfied (right).

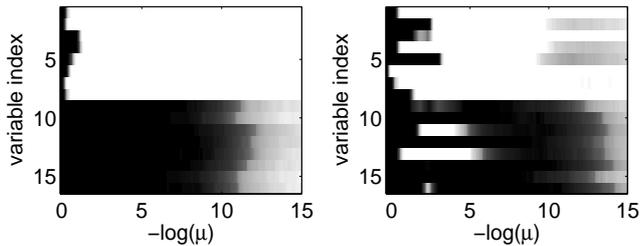


Figure 2. **Bolasso**: log-odd ratios of the probabilities of selection of each variable (white = large probabilities, black = small probabilities) vs. regularization parameter. Consistency condition in Eq. (2) satisfied (left) and not satisfied (right).

$p = 16$ and $r = 8$ relevant variables, one for which the consistency condition in Eq. (2) is satisfied (left), one for which it was not satisfied (right). For a fixed number of sample $n = 1000$, we generated 256 replications and computed the empirical frequencies of selecting any given variable for the Lasso as the regularization parameter μ varies. Those plots show the various asymptotic regimes of the Lasso detailed in Section 2. In particular, on the right plot, although no μ leads to perfect selection (i.e., exactly variables with indices less than $r = 8$ are selected), there is a range where all relevant variables are always selected, while all others are selected with probability within $(0, 1)$.

In Figure 2, we plot the results under the same conditions for the Bolasso (with a fixed number of bootstrap replications $m = 128$). We can see that in the Lasso-consistent case (left), the Bolasso widens the consistency region, while in the Lasso-inconsistent case (right), the Bolasso “creates” a consistency region.

In Figure 3, we selected the same two distributions and compared the probability of exactly selecting the correct support pattern, for the Lasso, and for the Bolasso with varying numbers of bootstrap replications (those probabilities are computed by averaging over 256 experiments with the same distribution). In Figure 3, we can see that in the Lasso-inconsistent case (right), the

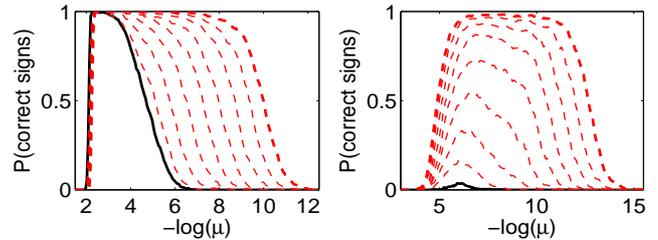


Figure 3. Bolasso (red, dashed) and Lasso (black, plain): probability of correct sign estimation vs. regularization parameter. Consistency condition in Eq. (2) satisfied (left) and not satisfied (right). The number of bootstrap replications m is in $\{2, 4, 8, 16, 32, 64, 128, 256\}$.

Bolasso indeed allows to fix the inability of the Lasso to find the correct pattern. Moreover, increasing m looks always beneficial; note that although it seems to contradict the asymptotic analysis in Section 3 (which imposes an upper bound for consistency), this is due to the fact that not selecting (at least) the relevant variables has very low probability and is not observed with only 256 replications.

Finally, in Figure 4, we compare various variable selection procedures for linear regression, to the Bolasso, with two distributions where $p = 64$, $r = 8$ and varying n . For all the methods we consider, there is a natural way to select exactly r variables with no free parameters (for the Bolasso, we select the most stable pattern with r elements, i.e., the pattern which corresponds to most values of μ). We can see that the Bolasso outperforms all other variable selection methods, even in settings where the number of samples becomes of the order of the number of variables, which requires additional theoretical analysis, subject of ongoing research. Note in particular that we compare with bagging of least-square regression (Breiman, 1996a) followed by a thresholding of the loading vector, which is another simple way of using bootstrap samples: the Bolasso provides a more efficient way to use the extra information, not for usual stabilization purposes (Breiman, 1996b), but directly for model selection. Note finally, that the bagging of Lasso estimates requires an additional parameter and is thus not tested.

4.2. UCI datasets

The previous simulations have shown that the Bolasso is successful at performing model selection in synthetic examples. We now apply it to several linear regression problems and compare it to alternative methods for linear regression, namely, ridge regression, Lasso, bagging of Lasso estimates (Breiman, 1996a), and a soft version of the Bolasso (referred to as Bolasso-S), where instead of intersecting the supports for each bootstrap

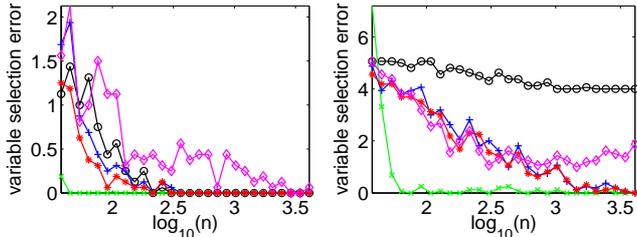


Figure 4. Comparison of several variable selection methods: Lasso (black circles), Bolasso (green crosses), forward greedy (magenta diamonds), thresholded LS estimate (red stars), adaptive Lasso (blue pluses). Consistency condition in Eq. (2) satisfied (left) and not satisfied (right). The averaged (over 32 replications) variable selection error is computed as the square distance between sparsity pattern indicator vectors.

replications, we select those which are present in at least 90% of the bootstrap replications. In Table 1, we consider data randomly generated as in Section 4.1 (with $p = 32$, $r = 8$, $n = 64$), where the true model is known to be composed of a sparse loading vector, while in Table 2, we consider regression datasets from the UCI machine learning repository. For all of those, we perform 10 replications of 10-fold cross validation and for all methods (which all have one free regularization parameter), we select the best regularization parameter on the 100 folds and plot the mean square prediction error and its standard deviation.

Note that when the generating model is actually sparse (Table 1), the Bolasso outperforms all other models, while in other cases (Table 2) the Bolasso is sometimes too strict in intersecting models, i.e., the softened version works better and is competitive with other methods. Studying the effects of this softened scheme (which is more similar to usual voting schemes), in particular in terms of the potential trade-off between good model selection and low prediction error, and under conditions where p is large, is the subject of ongoing work.

5. Conclusion

We have presented a detailed analysis of variable selection properties of a bootstrapped version of the Lasso. The model estimation procedure, referred to as the Bolasso, is provably consistent under general assumptions. This work brings to light that poor variable selection results of the Lasso may be easily enhanced thanks to a simple parameter-free resampling procedure. Our contribution also suggests that the use of bootstrap samples by L. Breiman in Bagging/Arcing/Random Forests (Breiman, 1998) may have been so far slightly overlooked and considered a

Table 1. Comparison of least-square estimation methods, data generated as described in Section 4.1, with $\kappa = \|\mathbf{Q}_{\mathbf{J}^c \mathbf{J}} \mathbf{Q}_{\mathbf{J} \mathbf{J}}^{-1} \mathbf{s}_{\mathbf{J}}\|_{\infty}$ (cf. Eq. (2)). Performance is measured through mean squared prediction error (multiplied by 100).

κ	0.93	1.20	1.42	1.28
Ridge	8.8 ± 4.5	4.9 ± 2.5	7.3 ± 3.9	8.1 ± 8.6
Lasso	7.6 ± 3.8	4.4 ± 2.3	4.7 ± 2.5	5.1 ± 6.5
Bolasso	5.4 ± 3.0	3.4 ± 2.4	3.4 ± 1.7	3.7 ± 10.2
Bagging	7.8 ± 4.7	4.6 ± 3.0	5.4 ± 4.1	5.8 ± 8.4
Bolasso-S	5.7 ± 3.8	3.0 ± 2.3	3.1 ± 2.8	3.2 ± 8.2

Table 2. Comparison of least-square estimation methods, UCI regression datasets. Performance is measured through mean squared prediction error (multiplied by 100).

	Autompg	Imports	Machine	Housing
Ridge	18.6 ± 4.9	7.7 ± 4.8	5.8 ± 18.6	28.0 ± 5.9
Lasso	18.6 ± 4.9	7.8 ± 5.2	5.8 ± 19.8	28.0 ± 5.7
Bolasso	18.1 ± 4.7	20.7 ± 9.8	4.6 ± 21.4	26.9 ± 2.5
Bagging	18.6 ± 5.0	8.0 ± 5.2	6.0 ± 18.9	28.1 ± 6.6
Bolasso-S	17.9 ± 5.0	8.2 ± 4.9	4.6 ± 19.9	26.8 ± 6.4

minor feature, while using bootstrap samples may actually be a key computational feature in such algorithms for good model selection performances, and eventually good prediction performances on real datasets.

The current work could be extended in various ways: first, we have focused on a fixed total number of variables, and allowing the numbers of variables to grow is important in theory and in practice (Meinshausen & Yu, 2006). Second, the same technique can be applied to similar settings than least-square regression with the ℓ_1 -norm, namely regularization by block ℓ_1 -norms (Bach, 2007) and other losses such as general convex classification losses. Finally, theoretical and practical connections could be made with other work on resampling methods and boosting (Bühlmann, 2006).

A. Proof of Model Consistency Results

In this appendix, we give sketches of proofs for the asymptotic results presented in Section 2 and Section 3. The proofs rely on the well-known property of the Lasso optimization problems, namely that if the sign pattern of the solution is known, then we can get the solution in closed form.

A.1. Optimality Conditions

We let denote $\bar{\varepsilon} = \bar{\mathbf{Y}} - \bar{\mathbf{X}} \mathbf{w} \in \mathbb{R}^n$, $\mathbf{Q} = \bar{\mathbf{X}}^{\top} \bar{\mathbf{X}} / n \in \mathbb{R}^{p \times p}$ and $\mathbf{q} = \bar{\mathbf{X}}^{\top} \bar{\varepsilon} / n \in \mathbb{R}^p$. First, we can equivalently rewrite Eq. (1) as:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2} (\mathbf{w} - \mathbf{w})^{\top} \mathbf{Q} (\mathbf{w} - \mathbf{w}) - \mathbf{q}^{\top} (\mathbf{w} - \mathbf{w}) + \mu_n \|\mathbf{w}\|_1. \quad (3)$$

The optimality conditions for Eq. (3) can be written in terms of the sign pattern $s = s(w) = \text{sign}(w)$ and the sparsity pattern $J = J(w) = \{j, w_j \neq 0\}$ (Yuan & Lin, 2007):

$$\begin{aligned} & \| (Q_{J^c J} Q_{J J}^{-1} Q_{J J} - Q_{J^c J}) \mathbf{w}_J + (Q_{J^c J} Q_{J J}^{-1} q_J - q_{J^c}) \\ & \quad + \mu_n Q_{J^c J} Q_{J J}^{-1} s_J \|_\infty \leq \mu_n, \end{aligned} \quad (4)$$

$$\text{sign}(Q_{J J}^{-1} Q_{J J} \mathbf{w}_J + Q_{J J}^{-1} q_J - \mu_n Q_{J J}^{-1} s_J) = s_J. \quad (5)$$

In this paper, we focus on regularization parameters μ_n of the form $\mu_n = \mu_0 n^{-1/2}$. The main idea behind the results is to consider that (Q, q) are distributed according to their limiting distributions, obtained from the law of large numbers and the central limit theorem, i.e., Q converges to \mathbf{Q} a.s. and $n^{1/2}q$ is asymptotically normally distributed with mean zero and covariance matrix $\sigma^2 \mathbf{Q}$. When assuming this, Propositions 1 and 2 are straightforward. The main effort is to make sure that we can safely replace (Q, q) by their limiting distributions. The following lemmas give sufficient conditions for correct estimation of the signs of variables in \mathbf{J} and for selecting a given pattern s (note that all constants could be expressed in terms of \mathbf{Q} and \mathbf{w} , details are omitted here):

Lemma 1 *Assume (A2) and $\|Q - \mathbf{Q}\|_2 \leq \lambda_{\min}(\mathbf{Q})/2$. Then $\text{sign}(\hat{w}_J) \neq \text{sign}(\mathbf{w}_J)$ implies $\|Q^{-1/2}q\|_2 \geq C_1 - \mu_n C_2$, where $C_1, C_2 > 0$.*

Lemma 2 *Assume (A2) and let $s \in \{-1, 0, 1\}^p$ such that $s_J = \text{sign}(\mathbf{w}_J)$. Let $J = \{j, s_j \neq 0\} \supset \mathbf{J}$. Assume*

$$\|Q - \mathbf{Q}\|_2 \leq \min\{\eta_1, \lambda_{\min}(\mathbf{Q})/2\}, \quad (6)$$

$$\|Q^{-1/2}q\|_2 \leq \min\{\eta_2, C_1 - \mu_n C_4\}, \quad (7)$$

$$\begin{aligned} & \|Q_{J^c J} Q_{J J}^{-1} q_J - q_{J^c} - \mu_n Q_{J^c J} Q_{J J}^{-1} s_J\|_\infty \leq \mu_n \\ & \quad - C_5 \eta_1 \mu_n - C_6 \eta_1 \eta_2, \end{aligned} \quad (8)$$

$$\forall i \in J \setminus \mathbf{J}, s_i [Q_{J J}^{-1}(q_J - \mu_n s_J)]_i \geq \mu_n C_7 \eta_1 + C_8 \eta_1 \eta_2, \quad (9)$$

with C_4, C_5, C_6, C_7, C_8 are positive constants. Then $\text{sign}(\hat{w}) = \text{sign}(\mathbf{w})$.

Those two lemmas are interesting because they relate optimality of certain sign patterns to quantities from which we can derive concentration inequalities.

A.2. Concentration Inequalities

Throughout the proofs, we need to provide upper bounds on the following quantities $\mathbb{P}(\|Q^{-1/2}q\|_2 > \alpha)$ and $\mathbb{P}(\|Q - \mathbf{Q}\|_2 > \eta)$. We obtain, following standard arguments (Boucheron et al., 2004): if $\alpha < C_9$ and $\eta < C_{10}$ (where $C_9, C_{10} > 0$ are constants),

$$\begin{aligned} \mathbb{P}(\|Q^{-1/2}q\|_2 > \alpha) & \leq 4p \exp\left(-\frac{n\alpha^2}{2pC_9}\right). \\ \mathbb{P}(\|Q - \mathbf{Q}\|_2 > \eta) & \leq 4p^2 \exp\left(-\frac{n\eta^2}{2p^2C_{10}}\right). \end{aligned}$$

We also consider multivariate *Berry-Esseen inequalities* (Bentkus, 2003); the probability $\mathbb{P}(n^{1/2}q \in \mathcal{C})$ can be estimated as $\mathbb{P}(t \in \mathcal{C})$ where t is normal with mean zero and covariance matrix $\sigma^2 \mathbf{Q}$. The error is then *uniformly* (for all convex sets \mathcal{C}) upper bounded by:

$$400p^{1/4}n^{-1/2}\lambda_{\min}(\mathbf{Q})^{-3/2}\mathbb{E}|\varepsilon|^3\|X\|_2^3 = C_{11}n^{-1/2}.$$

A.3. Proof of Proposition 1

By Lemma 2, for any given A , and n large enough, the probability that the sign is different from s is upper bounded by

$$\begin{aligned} & \mathbb{P}\left(\|Q^{-1/2}q\|_2 > \frac{A(\log n)^{1/2}}{n^{1/2}}\right) + \mathbb{P}\left(\|Q - \mathbf{Q}\|_2 > \frac{A(\log n)^{1/2}}{n^{1/2}}\right) \\ & \quad + \mathbb{P}\{t \notin \mathcal{C}(s, \mu_0(1 - \alpha))\} + 2C_{11}n^{-1/2}, \end{aligned}$$

where $\mathcal{C}(s, \beta)$ is the set of t such that (a) $\|Q_{J^c J} Q_{J J}^{-1} t_J - t_{J^c} - \beta Q_{J^c J} Q_{J J}^{-1} s_J\|_\infty \leq \beta$ and (b) for all $i \in J \setminus \mathbf{J}$, $s_i [Q_{J J}^{-1}(t_J - \beta s_J)]_i \geq 0$. Note that here $\alpha = O((\log n)n^{-1/2})$ tends to zero and that we have: $\mathbb{P}\{t \notin \mathcal{C}(s, \mu_0(1 - \alpha))\} \leq \mathbb{P}\{t \notin \mathcal{C}(s, \mu_0)\} + O(\alpha)$. All terms (if A is large enough) are thus $O((\log n)n^{-1/2})$.

This shows that $\mathbb{P}(\text{sign}(\hat{w}) = \text{sign}(\mathbf{w})) \geq \rho(s, \mu_0) + O((\log n)n^{-1/2})$ where $\rho(s, \mu_0) = \mathbb{P}\{t \in \mathcal{C}(s, \mu_0)\} \in (0, 1)$ —the probability is strictly between 0 and 1 because the set and its complement have non empty interiors and the normal distribution has a positive definite covariance matrix $\sigma^2 \mathbf{Q}$. The other inequality can be proved similarly. Note that the constant in $O((\log n)n^{-1/2})$ depends on μ_0 but by carefully considering this dependence on μ_0 , we could make the inequality uniform in μ_0 as long as μ_0 tends to zero or infinity at most at a logarithmic speed (i.e., μ_n deviates from $n^{-1/2}$ by at most a logarithmic factor). Also, it would be interesting to consider uniform bounds on portions of the regularization path.

A.4. Proof of Proposition 2

From Lemma 1, the probability of not selecting any of the variables in \mathbf{J} is upper bounded by $\mathbb{P}(\|Q^{-1/2}q\|_2 > C_1 - \mu_n C_2) + \mathbb{P}(\|Q - \mathbf{Q}\|_2 > \lambda_{\min}(\mathbf{Q})/2)$, which is straightforwardly upper bounded (using Section A.2) by a term of the required form.

A.5. Proof of Proposition 3

In order to simplify the proof, we made the simplifying assumption that the random variables X and ε have compact supports. Extending the proofs to take into account the looser condition that $\|X\|^2$ and ε^2 have non uniformly infinite cumulant generating functions (i.e., assumption (A1)) can be done with minor changes. The probability that $\bigcap_{k=1}^m J_k$ is different from \mathbf{J} is upper bounded by the sum of the following probabilities:

(a) Selecting at least variables in \mathbf{J} : the probability that for the k -th replication, one index in \mathbf{J} is not selected, each of them which is upper bounded by $\mathbb{P}(\|\mathbf{Q}^{-1/2}q^*\|_2 > C_1/2) + \mathbb{P}(\|\mathbf{Q} - \mathbf{Q}^*\|_2 > \lambda_{\min}(\mathbf{Q})/2)$, where q^* corresponds to the ghost sample; as common in theoretical analysis of the bootstrap, we relate q^* to q as follows: $\mathbb{P}(\|\mathbf{Q}^{-1/2}q^*\|_2 > C_1/2) \leq \mathbb{P}(\|\mathbf{Q}^{-1/2}(q^* - q)\|_2 > C_1/4) + \mathbb{P}(\|\mathbf{Q}^{-1/2}q\|_2 > C_1/4)$ (and similarly for $\mathbb{P}(\|\mathbf{Q} - \mathbf{Q}^*\|_2 > \lambda_{\min}(\mathbf{Q})/2)$). Because we have assumed that X and ε have compact supports, the bootstrapped variables have also compact support and we can use concentration inequalities (given the original variables \bar{X} , and also after expectation with respect to \bar{X}). Thus the probability for one bootstrap replication is upperbounded by Be^{-Cn} where B and C are strictly positive constants. Thus the overall contribution of this part is less than mBe^{-Cn} .

(b) Selecting at most variables in \mathbf{J} : the probability that for all replications, the set \mathbf{J} is not exactly selected (note that this is not tight at all since on top of the relevant variables which are selected with overwhelming probability, different additional variables may be selected for different replications and cancel out when intersecting).

Our goal is thus to bound $\mathbb{E}\{\mathbb{P}(\mathbf{J}^* \neq \mathbf{J}|\bar{X})^m\}$. By previous lemmas, we have that $\mathbb{P}(\mathbf{J}^* \neq \mathbf{J}|\bar{X})$ is upper bounded by $\mathbb{P}\left(\|\mathbf{Q}^{-1/2}q^*\|_2 > \frac{A(\log n)^{1/2}}{n^{1/2}}|\bar{X}\right) + \mathbb{P}\left(\|\mathbf{Q} - \mathbf{Q}^*\|_2 > \frac{A(\log n)^{1/2}}{n^{1/2}}|\bar{X}\right) + \mathbb{P}(t^* \notin \mathcal{C}(\mu_0)|\bar{X}) + 2C_{11}n^{-1/2} + O\left(\frac{\log n}{n^{1/2}}\right)$, where now, given \bar{X}, \bar{Y} , t^* is normally distributed with mean $n^{1/2}q$ and covariance matrix $\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 x_i x_i^\top$.

The first two terms and the last two ones are uniformly $O\left(\frac{\log n}{n^{1/2}}\right)$ (if A is large enough). We then have to consider the remaining term. We have $\mathcal{C}(\mu_0) = \{t^* \in \mathbb{R}^p, \|\mathbf{Q}_{\mathbf{J}^c \mathbf{J}} \mathbf{Q}_{\mathbf{J} \mathbf{J}}^{-1} t_{\mathbf{J}^c}^* - t_{\mathbf{J}^c}^* - \mu_0 \mathbf{Q}_{\mathbf{J}^c \mathbf{J}} \mathbf{Q}_{\mathbf{J} \mathbf{J}}^{-1} \mathbf{s}_{\mathbf{J}}\|_\infty \leq \mu_0\}$. By Hoeffding's inequality, we can replace the covariance matrix that depends on \bar{X} and \bar{Y} by $\sigma^2 \mathbf{Q}$, at cost $O(n^{-1/2})$. We thus have to bound $\mathbb{P}(n^{1/2}q + y \notin \mathcal{C}(\mu_0)|q)$ for y normally distributed and $\mathcal{C}(\mu_0)$ a fixed compact set. Because the set is compact, there exist constants $A, B > 0$ such that, if $\|n^{1/2}q\|_2 \leq \alpha$ for α large enough, then $\mathbb{P}(n^{1/2}q + y \notin \mathcal{C}(\mu_0)|q) \leq 1 - Ae^{-B\alpha^2}$. Thus, by truncation, we obtain a bound of the form: $\mathbb{E}\{\mathbb{P}(\mathbf{J}^* \neq \mathbf{J}|\bar{X})^m\} \leq (1 - Ae^{-B\alpha^2} + F \frac{\log n}{n^{1/2}})^m + Ce^{-B\alpha^2} \leq \exp(-mAe^{-B\alpha^2} + mF \frac{\log n}{n^{1/2}}) + Ce^{-B\alpha^2}$, where we have used Hoeffding's inequality to upper bound $\mathbb{P}(\|n^{1/2}q\|_2 > \alpha)$. By minimizing in closed form with respect to $e^{-B\alpha^2}$, i.e., with $e^{-B\alpha^2} = \frac{F \log n}{An^{1/2}} + \frac{\log(mA/C)}{mA}$, we obtain the desired inequality.

References

- Asuncion, A., & Newman, D. (2007). UCI machine learning repository.
- Bach, F. (2007). *Consistency of the group Lasso and multiple kernel learning* (Tech. report HAL-00164735). <http://hal.archives-ouvertes.fr>.
- Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113, 385–402.
- Boucheron, S., Lugosi, G., & Bousquet, O. (2004). Concentration inequalities. *Advanced Lectures on Machine Learning*. Springer.
- Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24, 123–140.
- Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Stat.*, 24, 2350–2383.
- Breiman, L. (1998). Arcing classifier. *Ann. Stat.*, 26, 801–849.
- Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Stat.*, 34, 559–583.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Ann. Stat.*, 32, 407.
- Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Chapman & Hall.
- Fu, W., & Knight, K. (2000). Asymptotics for Lasso-type estimators. *Ann. Stat.*, 28, 1356–1378.
- Meinshausen, N., & Yu, B. (2006). *Lasso-type recovery of sparse representations for high-dimensional data* (Tech. report 720). Dpt. of Statistics, UC Berkeley.
- Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, 58, 267–288.
- Wainwright, M. J. (2006). *Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1 -constrained quadratic programming* (Tech. report 709). Dpt. of Statistics, UC Berkeley.
- Yuan, M., & Lin, Y. (2007). On the non-negative garrotte estimator. *J. Roy. Stat. Soc. B*, 69, 143–161.
- Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *J. Mac. Learn. Res.*, 7, 2541–2563.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Stat. Ass.*, 101, 1418–1429.