Bolasso: Model Consistent Lasso Estimation through the Bootstrap

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure, Paris

July 2008

Outline

1. Review of asymptotic properties of the Lasso

2. Bolasso : using the bootstrap for consistent model selection

3. Simulations

Lasso

- Goal: predict a response $Y \in \mathbb{R}$ from $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ as a linear function $w^\top X$, with $w \in \mathbb{R}^p$

- Observations: *independent and identically distributed* (i.i.d.)
  - data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$
  - given in the form of matrices $\overline{Y} \in \mathbb{R}^n$ and $\overline{X} \in \mathbb{R}^{n \times p}$.

- Square loss function: $\frac{1}{2n} \sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2$

- Lasso: $$\boxed{\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2 + \mu_n \|w\|_1}$$

Lasso

- Goal: predict a response $Y \in \mathbb{R}$ from $X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ as a linear function $w^\top X$, with $w \in \mathbb{R}^p$

- Observations: *independent and identically distributed* (i.i.d.)
  - data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
  - given in the form of matrices $\overline{Y} \in \mathbb{R}^n$ and $\overline{X} \in \mathbb{R}^{n \times p}$.

- Square loss function: $\frac{1}{2n} \sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2$

- Lasso: $$\boxed{\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2 + \mu_n \|w\|_1}$$

- Regularization by $\|w\|_1$ leads to sparsity
  - Many efficient algorithms, empirical evaluations and extensions
  - Asymptotic analysis: does is actually work?

Asymptotic analysis

- Asymptotic set up
  - data generated from linear model $Y = X^\top \mathbf{w} + \varepsilon$
  - $\hat{w}$ any minimizer of the Lasso problem
  - number of observations n tends to infinity

- Three types of consistency
  - regular consistency: $\|\hat{w} - \mathbf{w}\|_2$ tends to zero in probability
  - pattern consistency: the sparsity pattern $\hat{J} = \{j,\ \hat{w}_j \neq 0\}$ tends to $\mathbf{J} = \{j,\ \mathbf{w}_j \neq 0\}$ in probability
  - sign consistency: the sign vector $\hat{s} = \mathrm{sign}(\hat{w})$ tends to $\mathbf{s} = \mathrm{sign}(\mathbf{w})$ in probability

- NB: with our assumptions, pattern and sign consistencies are equivalent once we have regular consistency

Assumptions for analysis

- Simplest assumptions (fixed p, large n):

1. Sparse linear model: $Y = X^\top \mathbf{w} + \varepsilon$, ε independent from X, and $\mathbf{w}$ sparse.
2. Finite cumulant generating functions $\mathbb{E}\exp(a\|X\|_2^2)$ and $\mathbb{E}\exp(a\varepsilon^2)$ finite for some $a > 0$.
3. Invertible matrix of second order moments $\mathbf{Q} = \mathbb{E}(XX^\top) \in \mathbb{R}^{p\times p}$.

Asymptotic analysis - simple cases

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2 + \mu_n \|w\|_1$$

- **If μ_n tends to infinity**
  - $\hat{w}$ tends to zero with probability tending to one
  - $\hat{J}$ tends to $\varnothing$ in probability

Asymptotic analysis - simple cases

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2 + \mu_n \|w\|_1$$

- **If μ_n tends to infinity**
  - $\hat{w}$ tends to zero with probability tending to one
  - $\hat{J}$ tends to $\varnothing$ in probability
- **If μ_n tends to $\mu_0 \in (0, \infty)$**
  - $\hat{w}$ converges to the minimum of $\frac{1}{2}(w - \mathbf{w})^\top \mathbf{Q}(w - \mathbf{w}) + \mu_0 \|w\|_1$
  - The sparsity and sign patterns may or may not be consistent
  - Possible to have sign consistency without regular consistency

Asymptotic analysis - simple cases

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\overline{Y} - \overline{X} w\|_2^2 + \mu_n \|w\|_1$$

- **If μ_n tends to infinity**
  - $\hat{w}$ tends to zero with probability tending to one
  - $\hat{J}$ tends to $\varnothing$ in probability
- **If μ_n tends to $\mu_0 \in (0, \infty)$**
  - $\hat{w}$ converges to the minimum of $\frac{1}{2}(w - \mathbf{w})^\top \mathbf{Q}(w - \mathbf{w}) + \mu_0 \|w\|_1$
  - The sparsity and sign patterns may or may not be consistent
  - Possible to have sign consistency without regular consistency
- **If μ_n tends to zero faster than $n^{-1/2}$**
  - $\hat{w}$ converges in probability to $\mathbf{w}$
  - With probability tending to one, all variables are included

Asymptotic analysis

$$\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|\overline{Y}-\overline{X}w\|_2^2+\mu_n\|w\|_1$$

- **If μ_n tends to zero slower than $n^{-1/2}$**
  - $\hat{w}$ converges in probability to $\mathbf{w}$
  - the sign pattern converges to the one of the minimum of

$$\frac{1}{2}v^\top\mathbf{Q}v+v_{\mathbf{J}}^\top\,\mathrm{sign}(\mathbf{w}_{\mathbf{J}})+\|v_{\mathbf{J}^c}\|_1$$

  - The sign pattern is equal to $\mathbf{s}$ (i.e., sign consistency) if and only if

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\,\mathrm{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty\leqslant 1$$

  - Consistency condition found by many authors: Yuan & Lin (2007), Wainwright (2006), Zhao & Yu (2007), Zou (2006)

Asymptotic analysis

$$\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|\overline{Y} - \overline{X}w\|_2^2 + \mu_n\|w\|_1$$

- **If μ_n tends to zero slower than $n^{-1/2}$**
  - $\hat{w}$ converges in probability to $\mathbf{w}$
  - the sign pattern converges to the one of the minimum of

$$\frac{1}{2}v^\top \mathbf{Q} v + v_{\mathbf{J}}^\top \operatorname{sign}(\mathbf{w}_{\mathbf{J}}) + \|v_{\mathbf{J}^c}\|_1$$

  - The sign pattern is equal to $\mathbf{s}$ (i.e., sign consistency) if and only if

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1} \operatorname{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1$$

  - Consistency condition found by many authors: Yuan & Lin (2007), Wainwright (2006), Zhao & Yu (2007), Zou (2006)
  - Disappointing?

Asymptotic analysis - new results

- **If μ_n tends to zero at rate $n^{-1/2}$, i.e., $n^{1/2}\mu_n \rightarrow \nu_0 \in (0, \infty)$**
  - $\hat{w}$ converges in probability to $\mathbf{w}$
  - All (and only) patterns which are consistent with $\mathbf{w}$ on $\mathbf{J}$ are attained with positive probability

Asymptotic analysis - new results

- **If μ_n tends to zero at rate $n^{-1/2}$, i.e., $n^{1/2}\mu_n \to \nu_0 \in (0,\infty)$**
  - $\hat{w}$ converges in probability to $\mathbf{w}$
  - All (and only) patterns which are consistent with $\mathbf{w}$ on $\mathbf{J}$ are attained with positive probability
  - **Proposition**: 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

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

  - **Proposition**: 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 $\rho(s,\nu_0) \in (0,1)$, and we have:

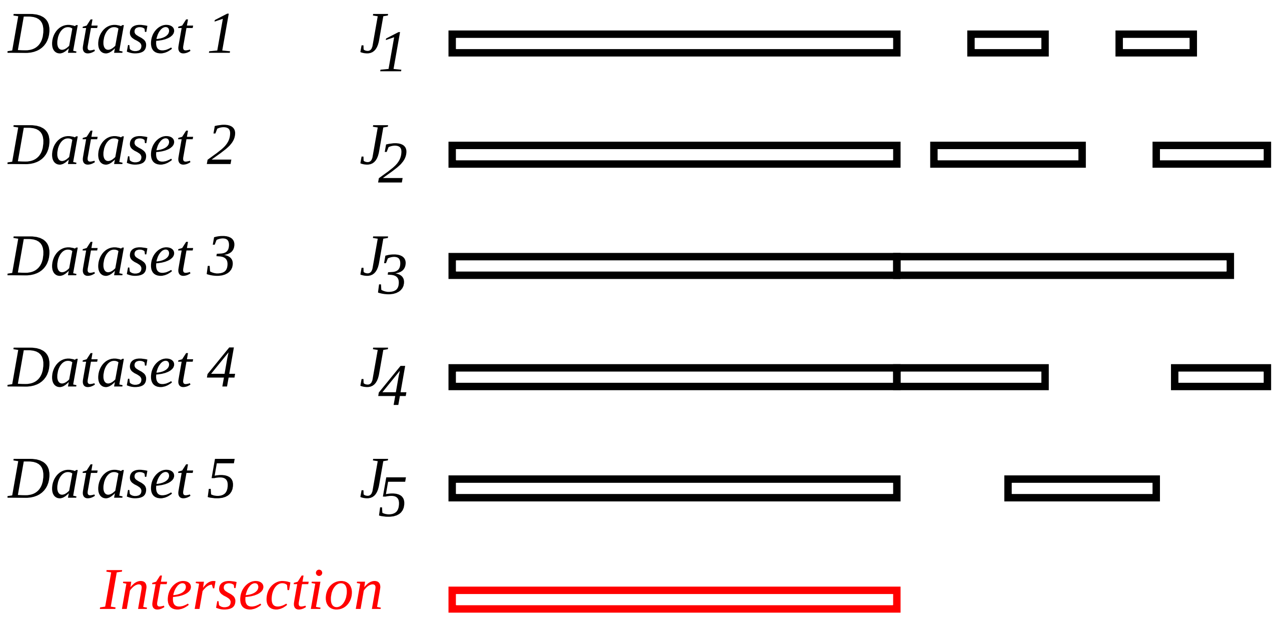
$$\mathbb{P}(\text{sign}(\hat{w}) = s) - \rho(s,\nu_0) = O(n^{-1/2}\log n).$$

μ_n tends to zero at rate $n^{-1/2}$

- Summary of asymptotic behavior:
  - All relevant variables (i.e., the ones in $\mathbf{J}$) are selected with probability tending to one exponentially fast
  - All other variables are selected with strictly positive probability

μ_n tends to zero at rate $n^{-1/2}$

- Summary of asymptotic behavior:
  - All relevant variables (i.e., the ones in $\mathbf{J}$) are selected with probability tending to one exponentially fast
  - All other variables are selected with strictly positive probability
- If several datasets (with same distributions) are available, intersecting support sets would lead to the correct pattern with high probability



Bootstrap

- Given n i.i.d. observations $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$

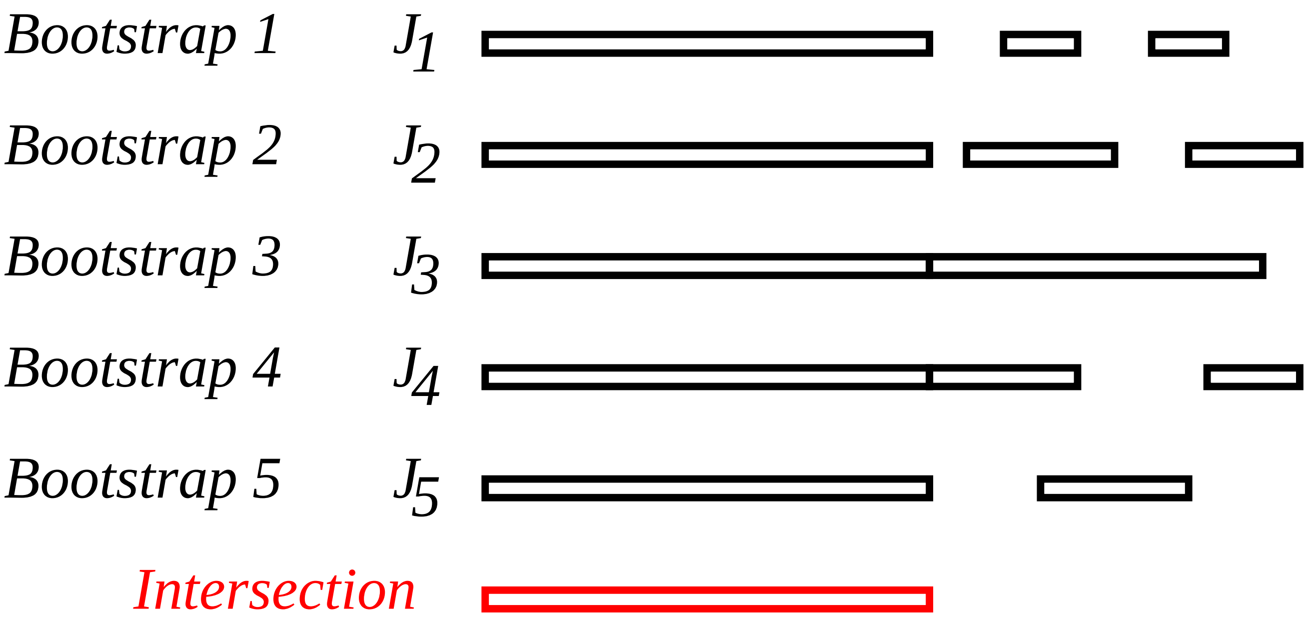
- m independent **bootstrap** replications: $k = 1, \ldots, m$,
  - *ghost samples* $(x_i^k, y_i^k) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, sampled independently and uniformly at random **with replacement** from the n original pairs

- Each bootstrap sample is composed of n potentially (and usually) duplicated copies of the original data pairs

- Standard way of mimicking availability of several datasets (Efron & Tibshirani, 1998)

Bolasso algorithm

- m applications of the Lasso/Lars algorithm (Efron et al., 2004)
  - Intersecting supports of variables
  - Final estimation of w on the entire dataset



Bolasso - Consistency result

- **Proposition**: Assume $\mu_n = \nu_0 n^{-1/2}$, with $\nu_0 > 0$. Then, for all $m > 1$, the probability that the Bolasso does not exactly select the correct model has the following upper bound:

$$\mathbb{P}(J \neq \mathbf{J}) \leqslant A_1 m e^{-A_2 n} + A_3 \frac{\log(n)}{n^{1/2}} + A_4 \frac{\log(m)}{m},$$

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

- Valid even if the Lasso consistency is not satisfied

- Influence of n, m

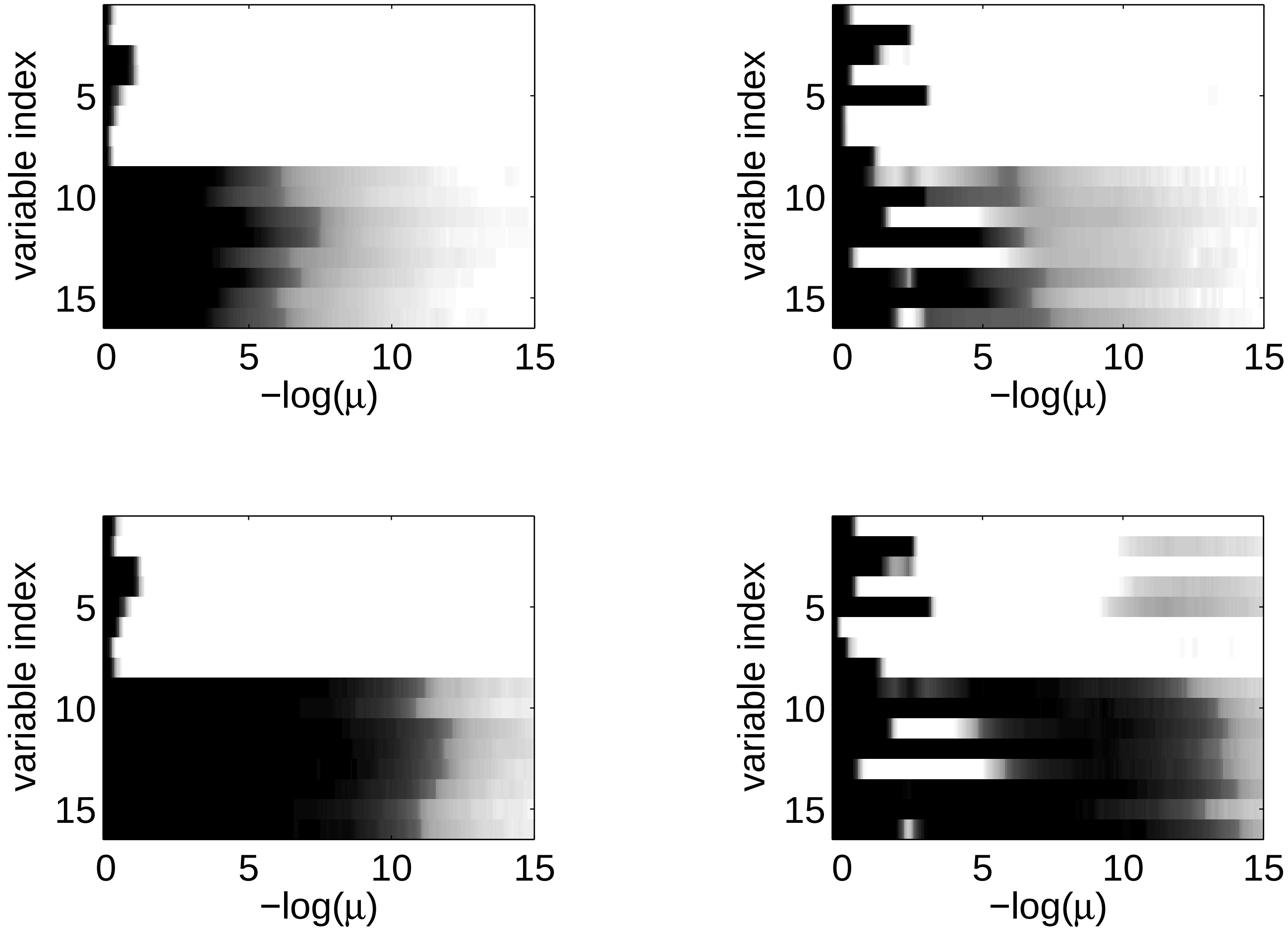
- Could be improved?

Consistency of the Lasso/Bolasso - Toy example

- Log-odd ratios of the probabilities of selection of each variable vs. μ

LASSO

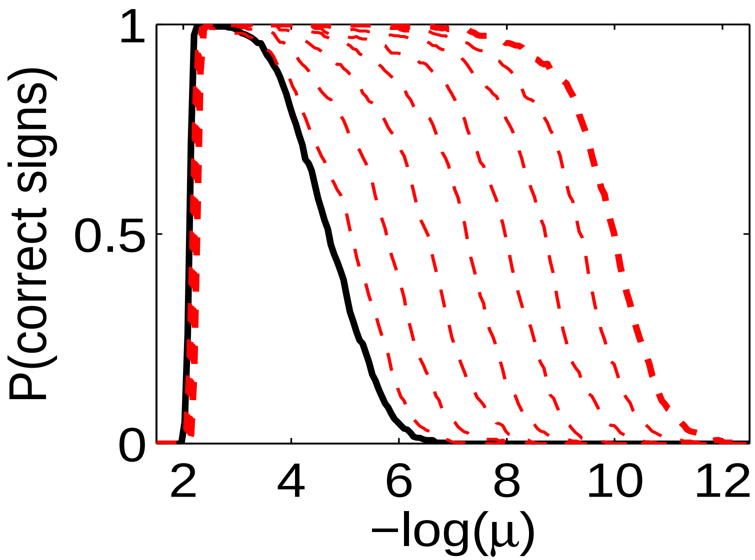
BOLASSO



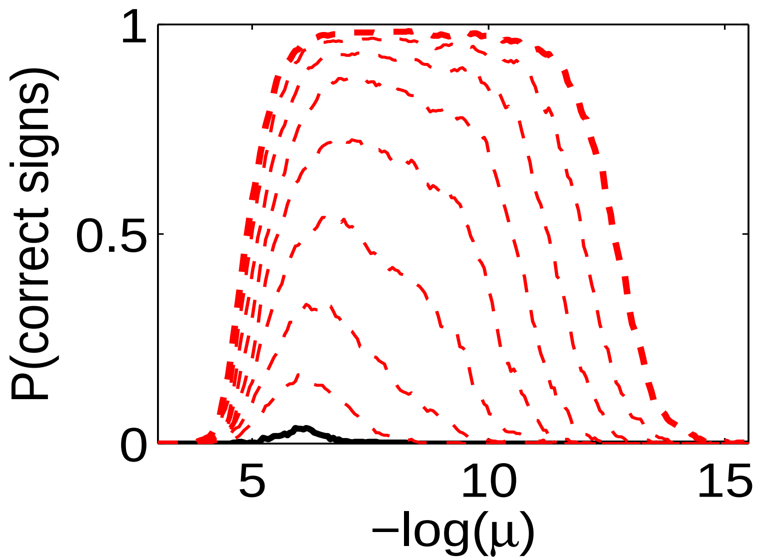
Consistency condition satisfied not satisfied

Influence of the number of bootstrap replications

- Bolasso (red) and Lasso (black): probability of correct sign estimation vs. regularization parameter, $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$.



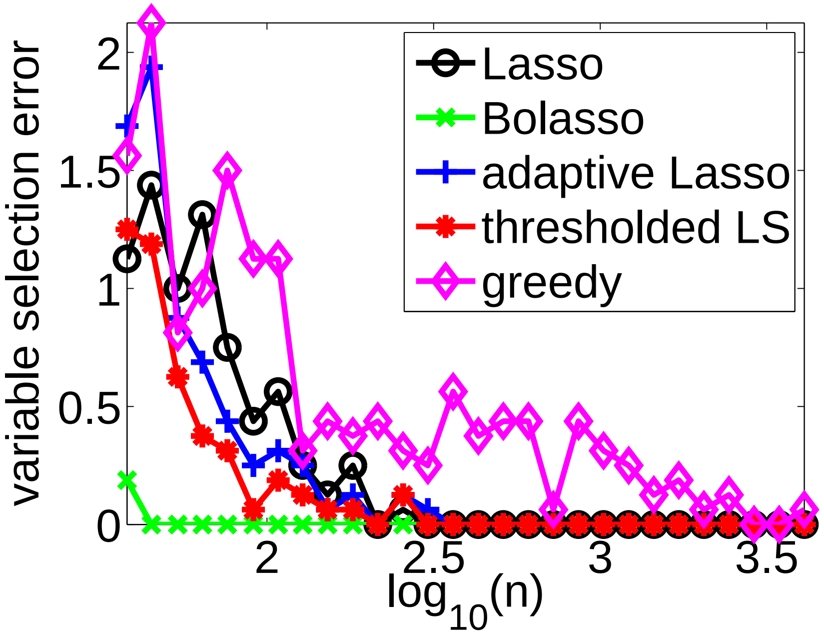
Consistency condition satisfied



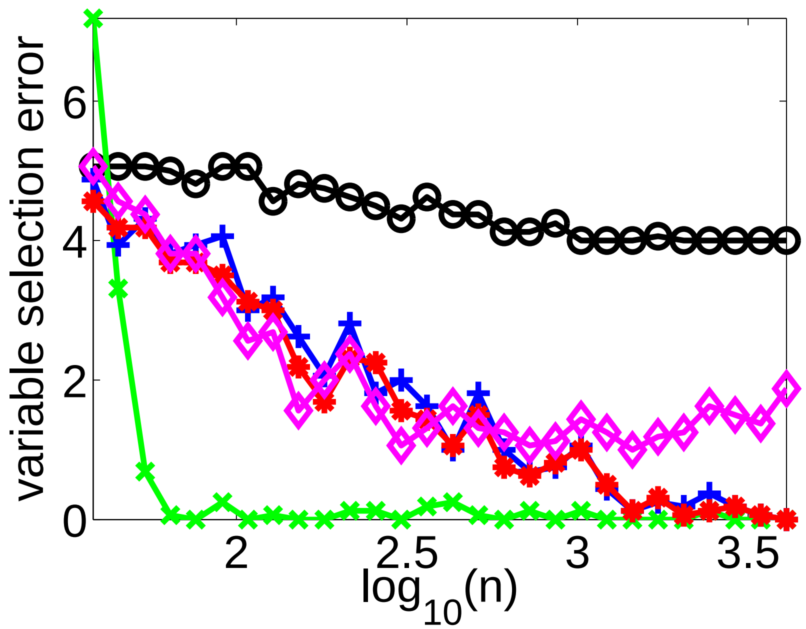
Consistency condition not satisfied

Comparison of several variable selection methods

- $p = 64$, averaged (over 32 replications) variable selection error = square distance between sparsity pattern indicator vectors.



Consistency condition satisfied



Consistency condition not satisfied

Comparison of least-square estimation methods

- Different values of $\kappa = \|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathbf{s}_{\mathbf{J}}\|_\infty$.

- Performance is measured through mean squared prediction error (multiplied by 100).

- Toy examples

- Regularization parameter estimated by cross-validation

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	$\mathbf{5.4 \pm 3.0}$	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	$\mathbf{3.0 \pm 2.3}$	$\mathbf{3.1 \pm 2.8}$	$\mathbf{3.2 \pm 8.2}$

Comparison of least-square estimation methods

- UCI regression datasets
- Performance is measured through mean squared prediction error (multiplied by 100).
- Regularization parameter estimated by cross-validation

	Autompg	Imports	Machine	Housing
Ridge	18.6 ± 4.9	$\mathbf{7.7} \pm \mathbf{4.8}$	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	$\mathbf{17.9} \pm \mathbf{5.0}$	8.2 ± 4.9	$\mathbf{4.6} \pm \mathbf{19.9}$	$\mathbf{26.8} \pm \mathbf{6.4}$

Conclusion

- Detailed analysis of variable selection properties of bootstrapped Lasso
- Consistency with no *consistency conditions* on covariance matrices
- No additional free parameter
- Extensions
  - Allowing p to grow (e.g., Meinshausen & Yu, 2008)
  - Extensions to the group Lasso (Yuan & Lin, 2006, Bach, 2008)
  - Connections with other resampling methods