Sparse Modeling: Lasso and Best Subset

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Tibshirani, Wasserman (2017). Sparsity, the Lasso, and Friends. Lecture notes on Statistical Machine Learning

Three norms: ℓ_0, ℓ_1 and ℓ_2

- Let's consider three canonical choices: the ℓ_0, ℓ_1 and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

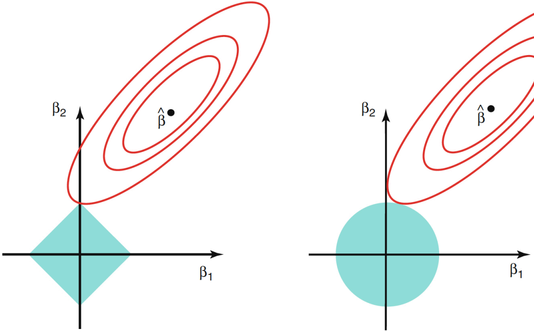
- ℓ_0 is not a proper norm: it does not satisfy positive homogeneity, i.e. $\|a\beta\|_0 \neq |a|\|\beta\|_0$ for $a \in \mathbb{R}$

Constrained form

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_0 \leq c \qquad \text{Best Subset Selection}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq c \qquad \text{Lasso Regression}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_2^2 \leq c \qquad \text{Ridge Regression}$$



The “classic” illustration comparing lasso and ridge constraints.
From Chapter 3 of ESL

Sparsity

- *Signal sparsity* is the assumption that only a small number of predictors have an effect, i.e. have $\beta_j \neq 0$
- In this case we would like our estimator $\hat{\beta}$ to be sparse, meaning that $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$
- Sparse estimators are desirable because perform variable selection and improve interpretability of the result
- The best subset selection and the lasso estimators are sparse, the ridge estimator is not sparse

Penalized form

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \qquad \text{Best Subset Selection}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{Lasso Regression}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad \text{Ridge Regression}$$

- Suppose that $y \sim N(\mu, 1)$
- ℓ_0 penalty

$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda \mathbb{1}\{\mu \neq 0\}, \qquad \hat{\mu} = H_{\sqrt{2\lambda}}(y)$$

  where $H_a(y) = y\mathbb{1}\{|y| > a\}$ is the hard-thresholding operator
- ℓ_1 penalty

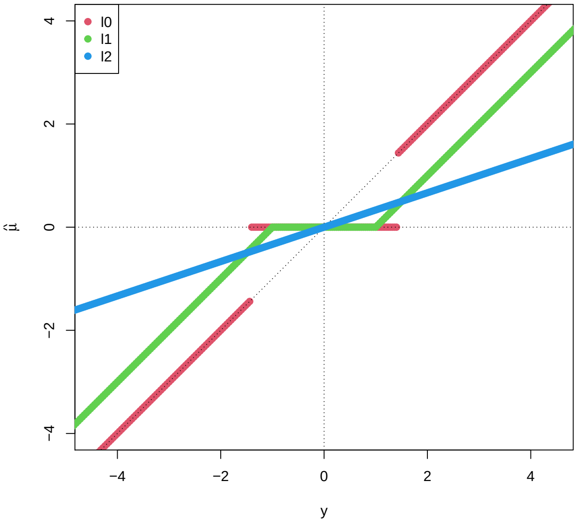
$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda|\mu|, \qquad \hat{\mu} = S_\lambda(y)$$

  where

$$S_a(y) = \begin{cases} y-a & \text{if } y > a \\ 0 & \text{if } -a \leq y \leq a \\ y+a & \text{if } y > a \end{cases}$$

  is the soft-thresholding operator
- ℓ_2 penalty

$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda\mu^2, \qquad \hat{\mu} = \left(\frac{1}{1+2\lambda}\right)y$$



$$\lambda = 1$$

Hard and soft thresholding

- ℓ_0 penalty creates a zone of sparsity but it is discontinuous (hard thresholding)
- ℓ_1 penalty creates a zone of sparsity but it is continuous (soft thresholding)
- ℓ_2 penalty creates a nice smooth estimator but it is never sparse

Orthogonal case

- Suppose $X^t X = I_p$
- OLS estimator

$$\hat{\beta}_j = X^t y$$

- BSS estimator

$$\hat{\beta}_j = H_{\sqrt{2\lambda}}(X^t y)$$

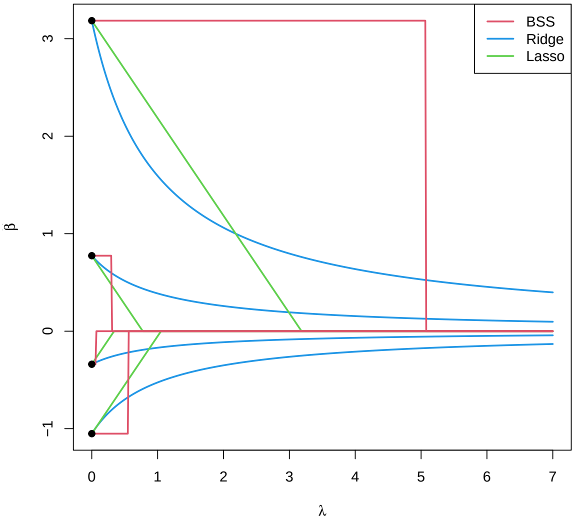
- Lasso estimator

$$\hat{\beta} = S_\lambda(X^t y)$$

- Ridge estimator

$$\hat{\beta} = (\frac{1}{1+2\lambda})X^t y$$

where $H_a(\cdot)$, $S_a(\cdot)$ are the componentwise hard- and soft-thresholding operators



Solution paths of ℓ_0, ℓ_1 and ℓ_2 penalties as a function of λ

Convexity

- Consider using the norm $\|\beta\|_q = (\sum_{j=1}^{q} |\beta_j|^q)^{1/q}$ as a penalty. Sparsity requires $q \leq 1$ and convexity requires $q \geq 1$. The only norm that gives sparsity and convexity is $q = 1$
- The lasso and ridge regression are *convex optimization problems*, best subset selection is not
- The ridge regression optimization problem is always *strictly convex* for $\lambda > 0$
- The best subset selection optimization problem is N-P-complete because of its combinatorial complexity (there are 2^p subsets), the worst kind of non convex problem

Best Subset Selection

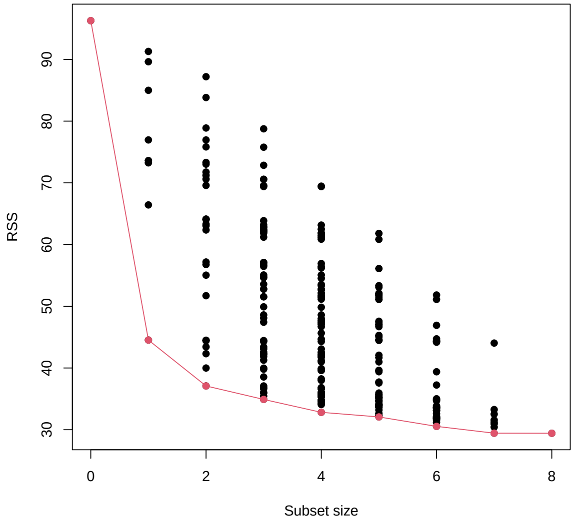
BSS algorithm

A natural approach is to consider all possible regression models each involving regressing the response on a different set of predictors

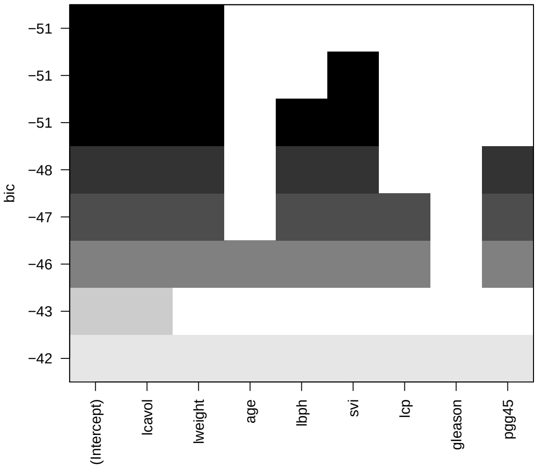
- Set B_0 as the null model (intercept only)
- For $k = 1, \ldots, p$
  1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
  2. Pick the best among these $\binom{p}{k}$ models, and call it B_k, where best is defined having the smallest residual sum of squares
- Select a single best model from among $B_0, B_1, \ldots, B_p$ (e.g. using Cp, BIC, Cross-Validation, validation set, etc.)

`prostate`

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	RSS
B_1	*								44.53
B_2	*	*							37.09
B_3	*	*			*				34.91
B_4	*	*		*	*				32.81
B_5	*	*		*	*			*	32.07
B_6	*	*		*	*	*		*	30.54
B_7	*	*	*	*	*	*		*	29.44
B_8	*	*	*	*	*	*	*	*	29.43



All possible subset models for the prostate cancer example



BIC Best Subset = B_2

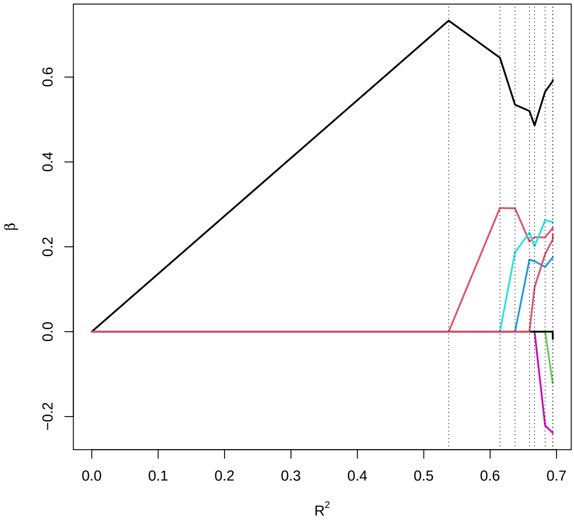
Computational bottleneck

- Furnival and Wilson (1974) and Hofmann et al. (2006) solve with $p \approx 30$ by using branch and bound algorithms. Implemented in the R packages `leaps` and `lmSubsets`
- Bertsimas et al. (2016) solve with $p \approx 100$ by using a mixed integer quadratic program along with the gurobi solver. Implemented in the R package `bestsubset`

Forward Stepwise Selection

Greedy forward algorithm, sub-optimal but feasible alternative to BSS and applicable when $p > n$

- Set S_0 as the null model (intercept only)
- For $k = 0, \ldots, \min(n-1, p-1)$:
    1. Consider all $p - k$ models that augment the predictors in S_k with one additional predictor
    2. Choose the best among these $p - k$ models and call it S_{k+1}, where best is defined having the smallest RSS
- Select a single best model from among $S_0, S_1, S_2, \ldots$ (e.g. using Cp, BIC, Cross-Validation, validation set, etc.)



Forward Stepwise solution path as a function of training R^2

The Lasso

- Lagrange form

$$\frac{1}{2n} \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- Intercept term omitted (center / scale y and the columns of X)
- The solution satisfies the subgradient / Karush-Kuhn-Tuker conditions

$$\frac{1}{n} X^t (y - X\hat{\beta}) = \lambda s$$

  where $s \in \partial \|\beta\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$

- The solution satisfies

$$-\frac{1}{n}\langle X_j, y - X\hat{\beta}\rangle + \lambda s_j = 0 \quad j = 1, \ldots, p$$

  where

$$s_j \in \begin{cases} 1 & \text{if } \hat{\beta}_j > 0 \\ [-1,1] & \text{if } \hat{\beta}_j = 0 \\ -1 & \text{if } \hat{\beta}_j < 0 \end{cases}$$

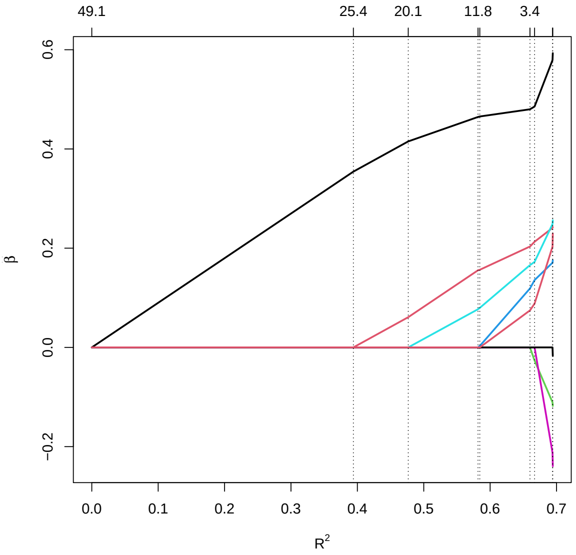
- Each of the variables in the model (with nonzero coefficient) has the same covariance with the residuals (in absolute value), i.e.

$$\frac{1}{n}|\langle X_j, y - X\hat{\beta}\rangle| = \lambda$$

- For all variables with zero coefficient

$$\frac{1}{n}|\langle X_j, y - X\hat{\beta}\rangle| \leq \lambda$$

- The coefficient profiles for the lasso are continuous and piecewise linear over the range of λ, with knots occurring whenever the *active set* changes, or the sign of the coefficients changes



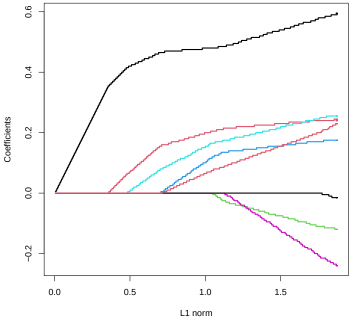
Lasso solution path as a function of training R^2

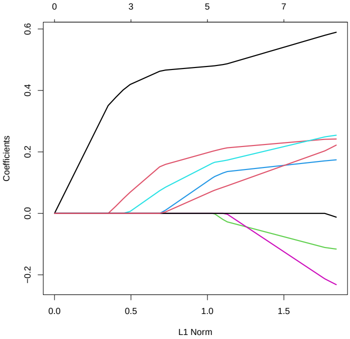
Boosting with componentwise linear least squares

- Response and predictors are standardized to have mean zero and unit norm
- Initialize $\hat{\beta}^{(0)} = 0$
- For $b = 1, \ldots, B$
  - compute the residuals $r = y - X\hat{\beta}^{(b-1)}$
  - find the predictor X_j most correlated with the residuals r
  - update $\hat{\beta}^{(b-1)}$ to $\hat{\beta}^{(b)}$ with

$$\hat{\beta}_j^{(b)} = \hat{\beta}_j^{(b-1)} + \epsilon \cdot s_j$$

    where s_j is the sign of the correlation
- This is known as *forward stagewise regression* and converges to the least squares solution when $n > p$
- Forward stagewise regression with infinitesimally small step-sizes, i.e. $\epsilon \to 0$, produces a set of solutions which is approximately equivalent to the set of Lasso solutions





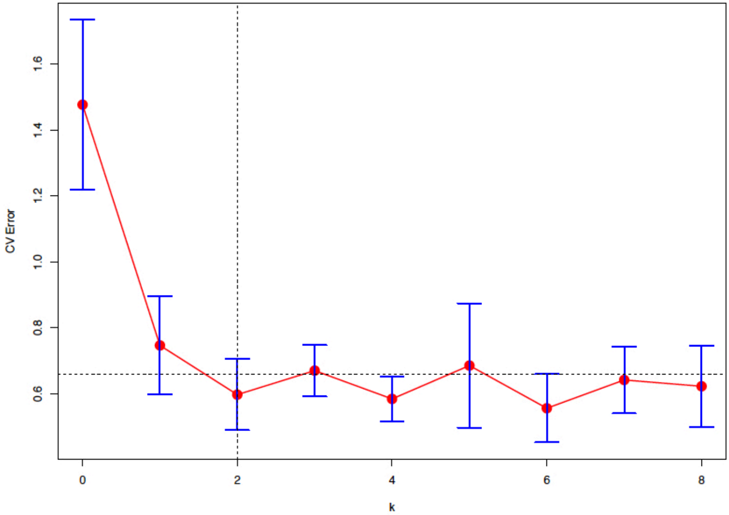
Left: forward stagewise regression with $\epsilon = 0.005$; Right: lasso

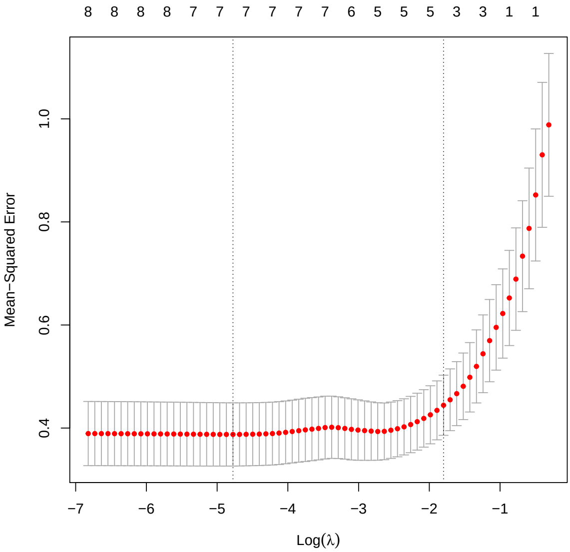
Cross-validation

- `lambda.min`: λ that minimize the cross-validation error
- `lambda.1se`: largest value of lambda such that error is within 1 standard error of the minimum (*one standard error rule*). To compute cross-validation ”standard errors”

  $$\mathrm{se} = \frac{1}{\sqrt{K}} \mathrm{sd}(\mathrm{Err}^{-1}, \ldots, \mathrm{Err}^{-K})$$

  where Err^{-k} denotes the error incurred in predicting the observations in the k hold-out fold, $k = 1, \ldots, K$.


CV Error
0.6
0.8
1.0
1.2
1.4
1.6
k
0
2
4
6
8



$\lambda_{\text{min}} = 0.008$ (7 nonzero), $\lambda_{\text{1se}} = 0.16$ (5 nonzero)

Extensions of the lasso