STK-IN4300
Statistical Learning Methods in Data Science

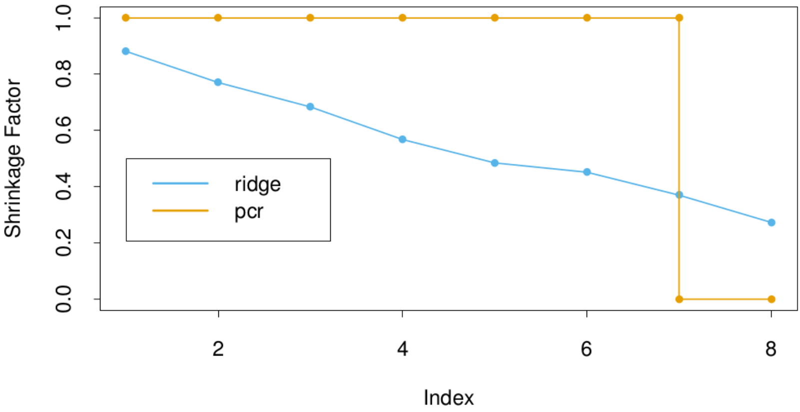
Riccardo De Bin

debin@math.uio.no

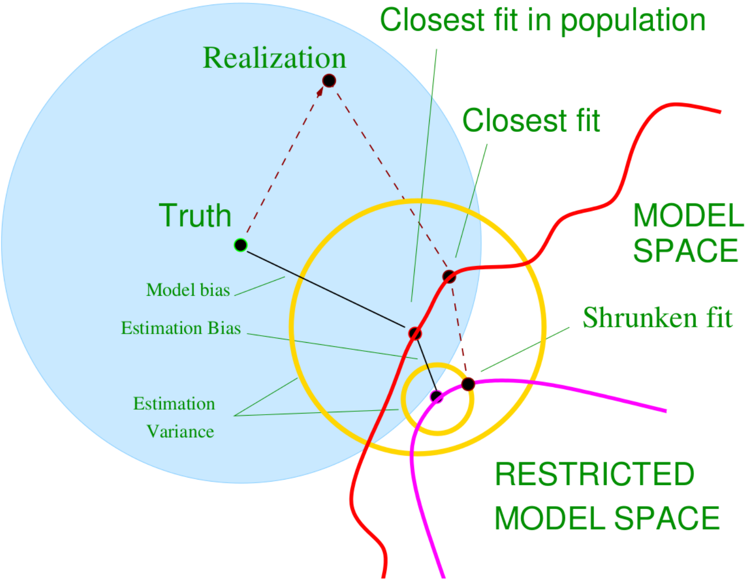
Outline of the lecture

- Shrinkage Methods
  - Lasso
  - Comparison of Shrinkage Methods
  - More on Lasso and Related Path Algorithms

Shrinkage Methods: ridge regression and PCR



Shrinkage Methods: bias and variance



Lasso: Least Absolute Shrinkage and Selection Operator

Lasso is similar to ridge regression, with an L_1 penalty instead of the L_2 one,

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2,$$

subject to $\sum_{j=1}^{p} |\beta_j| \leqslant t$.

Or, in the equivalent Lagrangian form,

$$\hat{\beta}_{\text{lasso}}(\lambda) = \text{argmin}_\beta \left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- X must be standardized;
- β_0 is again not considered in the penalty term.

Lasso: remarks

Due to the structure of the L_1 norm;

- some estimates are forced to be 0 (variable selection);
- no close form for the estimator.

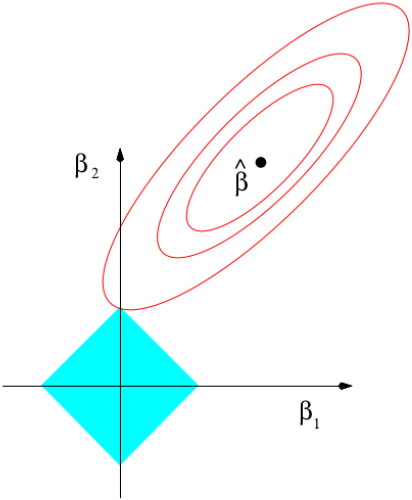
From a Bayesian prospective:

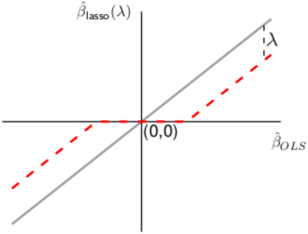
- $\hat{\beta}_{\text{lasso}}(\lambda)$ as the posterior mode estimate.
- $\beta \sim Laplace(0, \tau^2)$;
- for more details, see Park & Casella (2008).

Extreme situations:

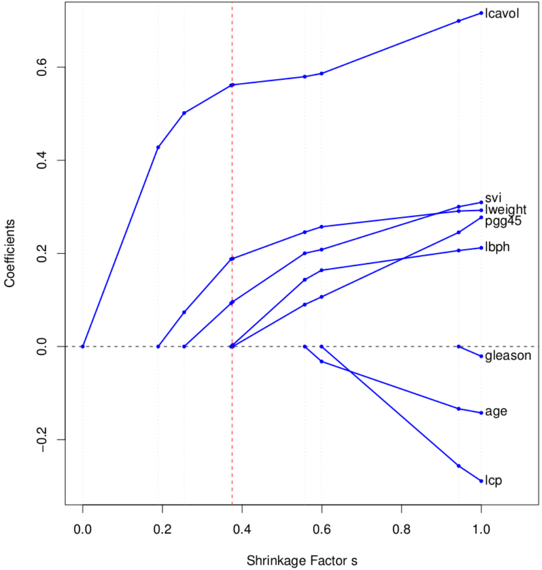
- $\lambda \to 0$, $\hat{\beta}_{\text{lasso}}(\lambda) \to \hat{\beta}_{\text{OLS}}$;
- $\lambda \to \infty$, $\hat{\beta}_{\text{lasso}}(\lambda) \to 0$.

Lasso: constrained estimation





Lasso: shrinkage



Lasso: generalized linear models

Lasso (and ridge r.) can be used with any linear regression model;

- e.g., logistic regression.

In logistic regression, the lasso solution is the maximizer of

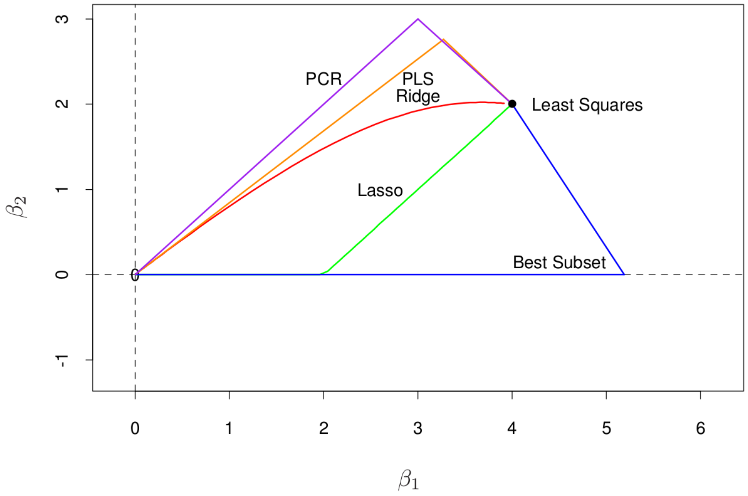
$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

Note:

- penalized logistic regression can be applied to problems with high-dimensional data (see Section 18.4).

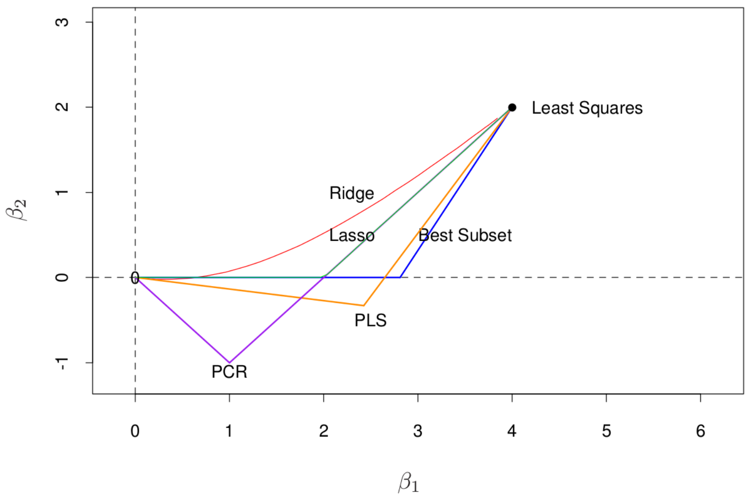
Comparison of Shrinkage Methods: coefficient profiles

$\rho = 0.5$

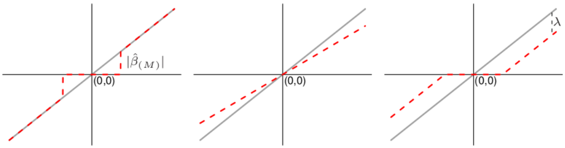


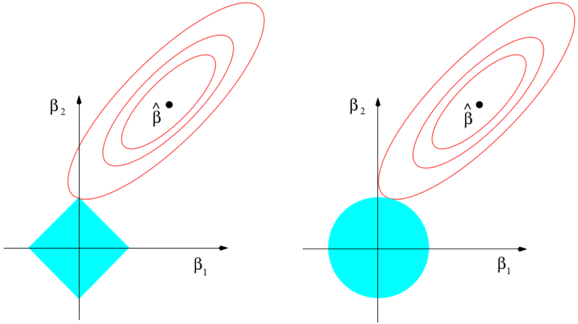
Comparison of Shrinkage Methods: coefficient profiles

$$\rho = -0.5$$



Comparison of Shrinkage Methods: coefficient profiles





More on Lasso and Related Path Algorithms: generalization

Generalization including lasso and ridge r. → **bridge regression**:

$$\tilde{\beta}(\lambda) = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, q \geqslant 0.$$

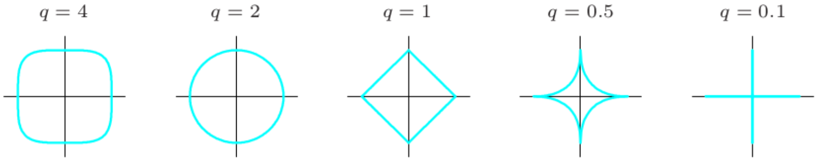
Where:

- $q = 0 \rightarrow$ best subset selection;
- $q = 1 \rightarrow$ lasso;
- $q = 2 \rightarrow$ ridge regression.

More on Lasso and Related Path Algorithms: generalization

Note that:

- $0 < q \leqslant 1 \rightarrow$ non differentiable;
- $1 < q < 2 \rightarrow$ compromise between lasso and ridge (but differentiable $\Rightarrow$ no variable selection property).
- q defines the shape of the constrain area:



- q could be estimated from the data (tuning parameter);
- in practice does not work well (variance).

More on Lasso and Related Path Algorithms: elastic net

Different compromise lasso / ridge regression: **elastic net**

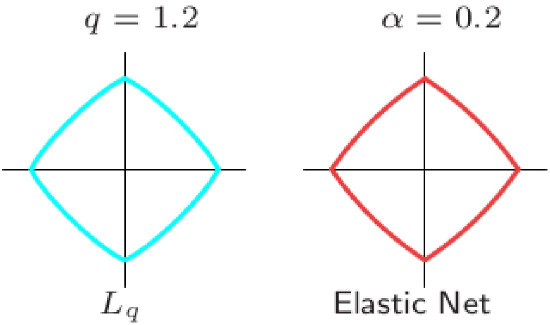
$$\tilde{\beta}(\lambda) = \text{argmin}_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right)\right\}.$$

Idea:

- L_1 penalty takes care of variable selection;
- L_2 penalty helps in correctly handling correlation;
- α defines how much L_1 and L_2 penalty should be used:
  - it is a tuning parameter, must be found in addition to λ;
  - a grid search is discouraged;
  - in real experiments, often very close to 0 or 1.

More on Lasso and Related Path Algorithms: elastic net

Comparing the bridge regression and the elastic net,



- they look very similar;
- huge difference due to differentiability (variable selection).

More on Lasso and Related Path Algorithms: Least Angle Regression

The Least Angle Regression (**LAR**):

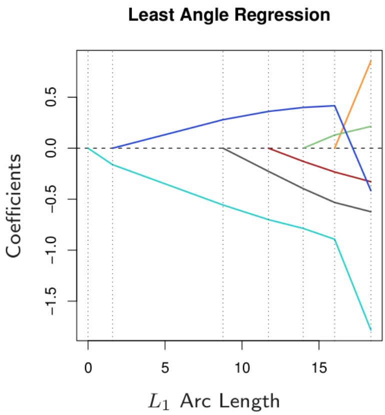
- can be viewed as a "democratic" version of the forward selection;
- add sequentially a new predictors into the model
  - only "as much as it deserves";
- eventually reaches the least square estimation;
- strongly connected with lasso;
  - lasso can be seen as a special case of LAR;
  - LAR is often used to fit lasso models.

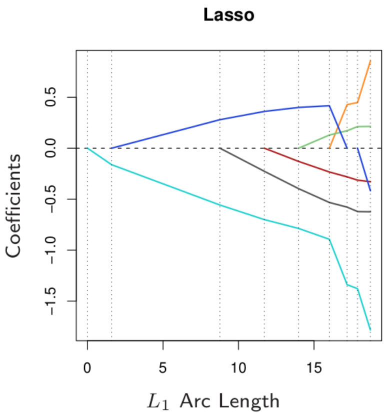
More on Lasso and Related Path Algorithms: LAR

Least Angle Regression:

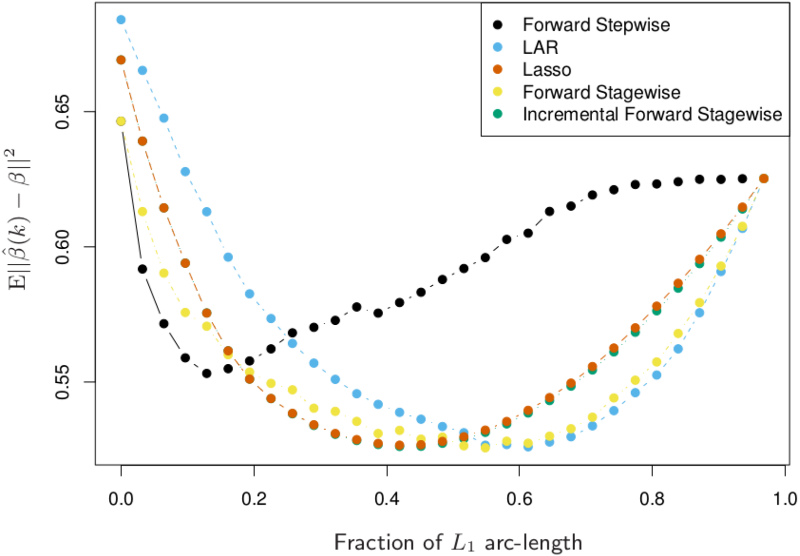
1. Standardize the predictors (mean zero, unit norm). Initialize:
   - residuals $r = y - \bar{y}$
   - regression coefficient estimates $\beta_1 = \cdots = \beta_p = 0$;
2. find the predictor x_j most correlated with r;
3. move $\hat{\beta}_j$ towards its least-squares coefficient $\langle x_j, r \rangle$,
   - until for $k \neq j$, corr$(x_k, r) =$ corr(x_j, r).
4. add x_k in the active list and update both $\hat{\beta}_j$ and $\hat{\beta}_k$:
   - towards their joint least squares coefficient;
   - until x_l has as much correlation with the current residual;
5. continue until all p predictors have been entered.

More on Lasso and Related Path Algorithms: comparison





More on Lasso and Related Path Algorithms: overfit



More on Lasso and Related Path Algorithms: other shrinkage methods

Group Lasso

Sometimes predictors belong to the same group:

- genes that belong to the same molecular pathway;
- dummy variables from the same categorical variable ...

Suppose the p predictors are grouped in L groups, group lasso minimizes

$$\min_\beta \left\{ ||(y - \beta_0 \vec{1} - \sum_{\ell=1}^{L} X_\ell \beta_\ell||_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\beta_j||_2 \right\},$$

where:

- $\sqrt{p_\ell}$ accounts for the group sizes;
- $||\cdot||$ denotes the (not squared) Euclidean norm
  - it is 0 $\Leftarrow$ all its component are 0;
- sparsity is encouraged at both group and individual levels.

More on Lasso and Related Path Algorithms: other shrinkage methods

Non-negative garrote

The idea of lasso originates from the non-negative garrote,

$$\hat{\beta}_{\text{garrote}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} c_j \beta_j x_{ij})^2,$$

subject to

$$c_j \geqslant 0 \text{ and } \sum_j c_j \leqslant t.$$

Non-negative garrote starts with OLS estimates and shrinks them:

- by non-negative factors;
- the sum of the non-negative factor is constrained;
- for more information, see Breiman (1995).

More on Lasso and Related Path Algorithms: other shrinkage methods

In the case of orthogonal design ($X^T X = I_N$),

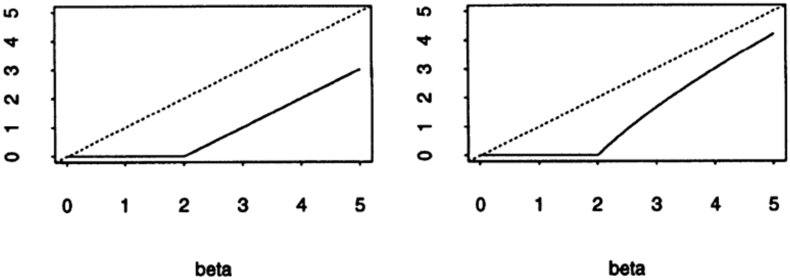
$$c_j(\lambda) = \left(1 - \frac{\lambda}{\hat{\beta}_j^{OLS}}\right),$$

where λ is a tuning parameter (related to t).

Note that the solution depends on $\hat{\beta}_{\text{OLS}}$:

- cannot be applied in $p >> N$ problems;
- may be a problem when $\hat{\beta}_{\text{OLS}}$ behaves poorly;
- has the oracle properties (Yuan & Lin, 2006) $\leftarrow$ see soon.

More on Lasso and Related Path Algorithms: other shrinkage methods



Comparison between lasso (left) and non-negative garrote (right).

(picture from Tibshirani, 1996)

More on Lasso and Related Path Algorithms: the oracle property

Let:

- $\mathcal{A} := \{j : \beta_j \neq 0\}$ be the set of the true relevant coefficients;
- δ be a fitting procedure (lasso, non-negative garrote, ...);
- $\hat{\beta}(\delta)$ the coefficient estimator of the procedure δ.

We would like that δ:

(a) identifies the right subset model, $\{j : \hat{\beta}(\delta) \neq 0\} = \mathcal{A}$;

(b) has the optimal estimation rate, $\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}}) \xrightarrow{d} N(0, \Sigma)$, where Σ is the covariance matrix for the true subset model.

If δ satisfies (a) and (b), it is called an oracle procedure.

More on Lasso and Related Path Algorithms: lasso and oracle property

Consider the following setup (Knight & Fu, 2000):

- $y_i = x_i\beta + \epsilon_i$, with ϵ_i i.i.d. r.v. with mean 0 and variance σ^2;
- $n^{-1}X^TX \to C$, where C is a positive definite matrix;
- suppose w.l.g. that $\mathcal{A} = \{1, 2, \dots, p_0\}$, $p_0 < p$;
- $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$, with C_{11} a $p_0 \times p_0$ matrix;
- $\hat{\mathcal{A}} = \{j : \hat{\beta}_j^{\text{lasso}}(\lambda) \neq 0\}$

More on Lasso and Related Path Algorithms: lasso and oracle property

Knight & Fu (2000) demonstrated the following two lemmas:

Lemma (1)

If $\lambda/n \to \lambda_0 \geqslant 0$, *then* $\hat{\beta}^{lasso}(\lambda) \xrightarrow{p} \text{argmin}_u V_1(u)$, *where*

$$V_1(u) = (u - \beta)^T C (u - \beta)^T + \lambda_0 \sum_{i=1}^{p} |u_j|.$$

Lemma (2)

If $\lambda/\sqrt{n} \to \lambda_0 \geqslant 0$, *then* $\sqrt{n}(\hat{\beta}^{lasso}(\lambda) - \beta) \xrightarrow{d} \text{argmin}_u V_2(u)$,

$$V_2(u) = -2u^T W + u^T C u + \lambda_0 \sum_{i=1}^{p} [u_j \, sgn(\beta) \mathbb{1}(\beta \neq 0) + |u_j| \mathbb{1}(\beta = 0)].$$

More on Lasso and Related Path Algorithms: lasso and oracle property

From Lemma (1):

- only $\lambda_0 = 0$ guarantees consistency.

Lemma (2) states:

- the lasso estimate is $\sqrt{n}$-consistent;
- when $\lambda = O(\sqrt{n})$, $\hat{\mathcal{A}}$ cannot be $\mathcal{A}$ with positive probability.

Proposition (1)

If $\lambda/\sqrt{n} \to \lambda_0 \geqslant 0$, *then lim sup*$_n P[\hat{\mathcal{A}} = \mathcal{A}] \leqslant c < 1$.

For the proof, see Zou (2006).

More on Lasso and Related Path Algorithms: lasso and oracle property

It may be interesting to see what happens in the intermediate case, when $\lambda_0 = \infty$, i.e., $\lambda/n \to 0$ and $\lambda/\sqrt{n} \to \infty$.

Lemma (3)

If $\frac{\lambda}{n} \to 0$ and $\frac{\lambda}{\sqrt{n}} \to \infty$, then $\frac{n}{\lambda}(\hat{\beta}^{lasso}(\lambda) - \beta) \xrightarrow{p} \text{argmin}_u V_3(u)$,

$$V_3(u) = u^T C u + \sum_{i=1}^{p} [u_j \mathit{sgn}(\beta) \mathbb{1}(\beta \neq 0) + |u_j| \mathbb{1}(\beta = 0)].$$

Note:

- the convergence rate of $\hat{\beta}^{\text{lasso}}(\lambda)$ is slower than $\sqrt{n}$;
- the optimal estimation rate is available only when $\lambda = O(\sqrt{n})$, but it leads to inconsistent variable selection;
- for the proof, see Zou (2006).

More on Lasso and Related Path Algorithms: necessary condition

Can consistency in variable selection can be achieved by sacrificing the rate of convergence in estimation?

$\downarrow$

Non necessarily.

It is possible to derive a necessary condition for consistency of the lasso variable selection (Zou, 2006):

Theorem (necessary condition)

Suppose that $\lim_n P[\hat{\mathcal{A}} = \mathcal{A}] = 1$. *Then there exists some sign vector* $s = (s_1, \dots, s_{p_0})^T$, $s_j \in \{-1, 1\}$, *such that*

$$|C_{21} C_{11}^{-1} s| \leqslant 1. \tag{1}$$

The last equation is understood componentwise.

More on Lasso and Related Path Algorithms: necessary condition

If condition (1) fails $\Rightarrow$ the lasso variable selection is inconsistent.

Corollary (1)

Suppose that $p_0 = 2m + 1 \geqslant 3$ *and* $p = p_0 + 1$, *so there is one irrelevant predictor. Let* $C_{11} = (1 - \rho_1)I + \rho_1 J_1$, *where* J_1 *is the matrix of 1's,* $C_{12} = \rho_2 \vec{1}$ *and* $C_{22} = 1$. *If* $-\frac{1}{p_0 - 1} < \rho_1 < \frac{1}{p_0}$ *and* $1 + (p_0 - 1\rho_1) < |\rho_2| < \sqrt{(1 + (p_0 - 1)/\rho_1/p_0)}$, *then condition* (1) *cannot be satisfied. So the lasso variable selection is inconsistent.*

More on Lasso and Related Path Algorithms: Corollary (1)

Proof of Corollary (1).

Note that

- $C_{11}^{-1} = \frac{1}{1-\rho_1}(I - \frac{\rho_1}{1+(p_0-1)\rho_1} J_1)$;
- $C_{21}C_{11}^{-1} = \frac{\rho_2}{1+(p_0-1)\rho_1}(\vec{1})^T$.

Therefore $C_{21}C_{11}^{-1}s = \frac{\rho_2}{1+(p_0-1)\rho_1}(\sum_{j=1}^{p_0} s_j)\vec{1}$.

Then, condition (1) becomes $\left|\frac{\rho_2}{1+(p_0-1)\rho_1}\right| \cdot \left|\sum_{j=1}^{p_0} s_j\right| \leqslant 1$.

Note that when p_0 is a odd number, $|\sum_{j=1}^{p_0} s_j| \geqslant 1$.

If $|\frac{\rho_2}{1+(p_0-1)\rho_1}| > 1$, then condition (1) cannot be satisfied for any sign vector. The choice of (ρ_1, ρ_2) ensures that C is a positive matrix and $|\frac{\rho_2}{1+(p_0-1)\rho_1}| > 1$. $\square$

More on Lasso and Related Path Algorithms: other shrinkage methods

The Smoothly Clipped Absolute Deviation (**SCAD**) estimator

$$\hat{\beta}_{\text{scad}}(\lambda, \alpha) = \text{argmin}_{\beta} \left\{ \frac{1}{2} ||y - \beta_0 \vec{1} - X\beta||_2^2 + \lambda \sum_{j=1}^{p} p_j(\beta_j; \lambda, \alpha) \right\},$$

where

$$\frac{d\, p_j(\beta_j; \lambda, \alpha)}{d\beta_j} = \lambda \left\{ \mathbb{1}(|\beta_j| \leqslant \lambda) + \frac{(\alpha\lambda - |\beta_j|)_+}{(\alpha - 1)\lambda} \mathbb{1}(|\beta_j| > \lambda) \right\}$$

for $\alpha > 2$.

Usually:

- α is set equal to 3.7 (based on simulations);
- λ is chosen via cross-validation.

More on Lasso and Related Path Algorithms: other shrinkage methods

The SCAD penalty function:

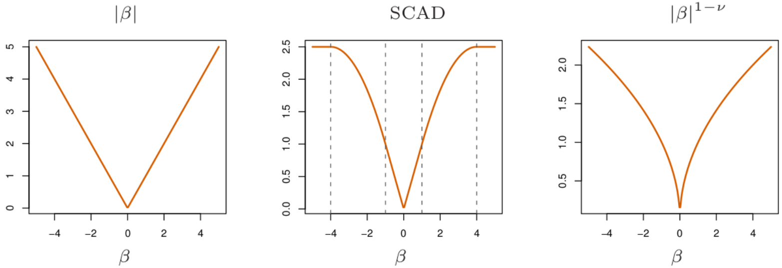
- penalizes less the largest regression coefficient estimates;
- makes the solution continuous. In particular,

$$\hat{\beta}_{\text{scad}}(\lambda, \alpha) = \begin{cases} \text{sgn}(\beta)(|\beta_j| - \lambda)_+ & \text{when } |\beta| \leqslant 2\lambda \\ \{(\alpha - 1)\beta - \text{sgn}(\beta)\alpha\lambda\}/(\alpha - 2) & \text{when } 2\lambda < |\beta| \leqslant \alpha\lambda \\ \beta & \text{when } |\beta| > \alpha\lambda \end{cases}$$

Note that:

- $\exists$ an $\sqrt{n}$-consistent estimator (Fan & Li, 2001, Theorem 1);
- the SCAD estimator $\hat{\beta}_{\text{scad}}(\lambda, \alpha)$ is an oracle estimator (Fan & Li, 2001, Theorem 2).

More on Lasso and Related Path Algorithms: other shrinkage methods



More on Lasso and Related Path Algorithms: other shrinkage methods

Adaptive lasso

The adaptive lasso is a particular case of the weighted lasso,

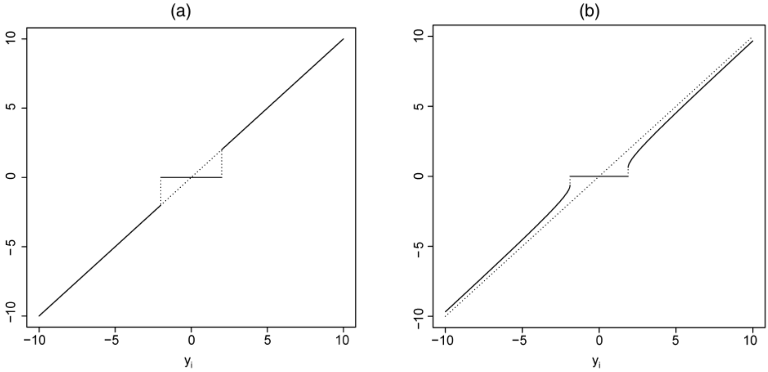
$$\hat{\beta}_{\text{weight}}(\lambda) = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\},$$

in which $\hat{w}_j = 1/|\hat{\beta}_{\text{OLS}}|^{\gamma}$.

Note:

- it enjoys the oracle properties (Zou, 2006, Theorem 2);
- when $\gamma = 1$, it is very closely related to the non-negative garrote (there is an additional sign constrain);
- relies on $\hat{\beta}_{\text{OLS}}$ $\rightarrow$ sometimes lasso used in a first step;
- two-dimensional tuning parameter.

More on Lasso and Related Path Algorithms: other shrinkage methods

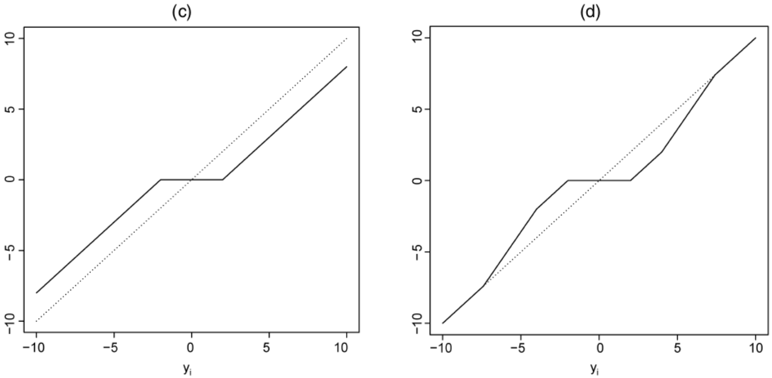


(a) best subset regression;

(b) bridge with $\alpha = 0.5$.

(picture from Zou, 2006)

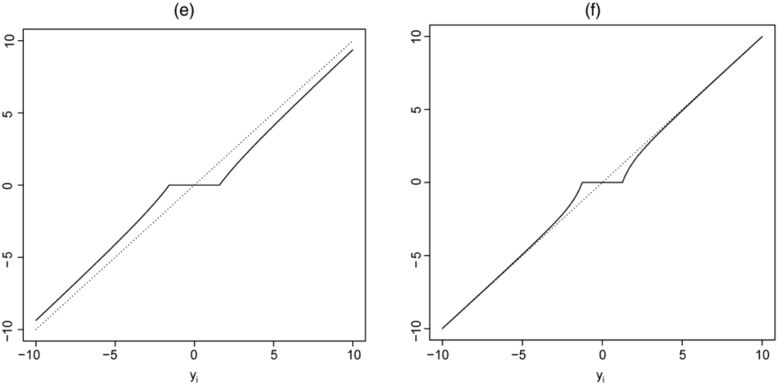
More on Lasso and Related Path Algorithms: other shrinkage methods



- (c) lasso;
- (d) scad.

(picture from Zou, 2006)

More on Lasso and Related Path Algorithms: other shrinkage methods



(e) adaptive lasso with $\gamma = 0.5$;

(f) adaptive lasso with $\gamma = 0.2$.

(picture from Zou, 2006)

References I

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

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

KNIGHT, K. & FU, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.

PARK, T. & CASELLA, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

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

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

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