A Short Introduction to the Lasso Methodology

Michael Gutmann

sites.google.com/site/michaelgutmann

University of Helsinki   Aalto University

Helsinki Institute for Information Technology

March 9, 2016

Goals of the lecture

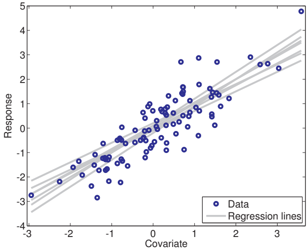
Lasso $\equiv$ Least Absolute Shrinkage and Selection Operator

Goal: After the lecture, to understand what these words mean

- *Shrinkage*: The lasso shrinks / regularizes the least squares regression coefficients (like ridge regression).
- *Selection*: The lasso also performs variable selection (unlike ridge regression).
- *Least absolute*: Shrinkage and selection are achieved by penalizing the absolute values of the regression coefficients.

Linear regression

- Data: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
  - n observations of pairs $(\mathbf{x}_i, y_i)$
  - $\mathbf{x}_i \in \mathbb{R}^p$: covariates
  - $y_i \in \mathbb{R}$: response



- Assumption: linear relation between covariates and response

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i \tag{1}$$
$$= \mathbf{x}_i^\top \boldsymbol{\beta} + e_i \tag{2}$$

where e_i is the residual

- Goal: Determine the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$

Least squares

- Minimize the residual sum of squares (RSS)

$$\text{RSS}(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \beta \right)^2 \tag{3}$$

- In vector notation, with

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \tag{4}$$

we have

$$\text{RSS}(\beta) = ||\boldsymbol{y} - \mathbf{X}\beta||_2^2 \tag{5}$$

Least squares

- Closed form solution

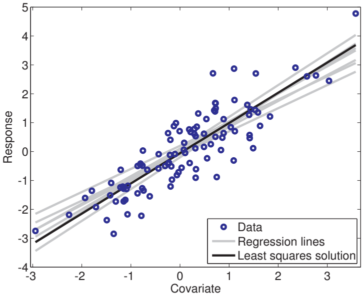
$$\hat{\beta}^o = \underset{\beta}{\operatorname{argmin}} \operatorname{RSS}(\beta) \quad (6)$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y} \quad (7)$$

if $p \times p$ matrix $\mathbf{X}^\top \mathbf{X}$ is invertible

- Prediction given a test covariate vector $\boldsymbol{x}$

$$\hat{y} = \boldsymbol{x}^\top \hat{\beta}^o \quad (8)$$



Ridge regression

- If $\mathbf{X}^\top\mathbf{X}$ is not invertible, regularized inverse can be taken

$$\hat{\beta}^r = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{y} \tag{9}$$

  where $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\lambda \geq 0$ the regularization parameter.
- This is ridge regression, $\hat{\beta}^r$ is minimizing $J^r(\beta)$

$$J^r(\beta) = ||\boldsymbol{y} - \mathbf{X}\beta||_2^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{10}$$

- As λ increases, $\hat{\beta}^r$ shrinks to zero ("shrinkage").

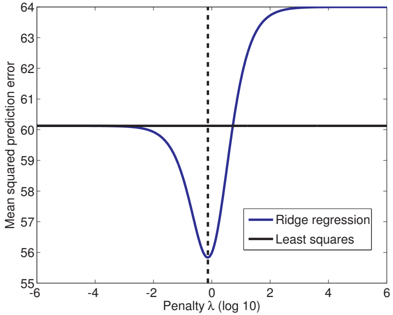
Benefits of ridge regression

$$\hat{\beta}^r = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- Regularization / shrinkage is useful even if $\mathbf{X}^\top \mathbf{X}$ is invertible.
- Reason: it can improve prediction accuracy

Example:

- $n = 50$ observations, $p = 10$ covariates
- Orthonormal matrix $\mathbf{X}$: $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$
- $\hat{\beta}^r = \frac{1}{1+\lambda} \mathbf{X}^\top \boldsymbol{y} = \frac{1}{1+\lambda} \hat{\beta}^o$

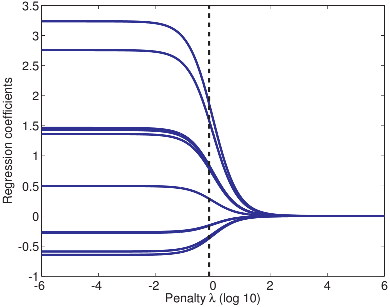


Limits of ridge regression

$\hat{\beta}^r = \frac{1}{1+\lambda}\hat{\beta}^o$

- Data were artificially generated with $\boldsymbol{\beta}^* = (3, 2, 1, 0, \ldots, 0)^\top$
- The vector is sparse: only 3/10 nonzero terms
- Ridge regression cannot recover sparse $\boldsymbol{\beta}$.

- Ridge regression performs shrinkage but not variable selection.
- *Variable selection*: Some $\hat{\beta}_j$ are set to zero; covariates are omitted from the fitted model.



Some practical aspects

- Choice of λ: via cross-validation
- Ridge solution $\hat{\beta}^r$ depends on the scale of the covariates.
  - $\rightarrow$ Center so that $\sum_{i=1}^n y_i = \sum_{i=1}^n x_{ij} = 0$
  - $\rightarrow$ Re-scale so that $\sum_{i=1}^n x_{ij}^2 = 1$
- Assume that the data were preprocessed in this manner.

Importance of variable selection

- It reduces the complexity of the models.
- The models become easier to interpret.
- It makes prediction cheaper: only covariates with nonzero $\hat{\beta}_j$ need to be measured.

$$\hat{y} = x_1\hat{\beta}_1 + \ldots + x_{1000}\hat{\beta}_{1000}$$

$$\downarrow$$

$$\hat{y} = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + x_3\hat{\beta}_3$$

Lasso regression

- Lasso regression consists in minimizing $J^L(\beta)$,

$$J^L(\beta) = ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

- Similar to the cost function $J^r(\beta)$ for ridge regression,

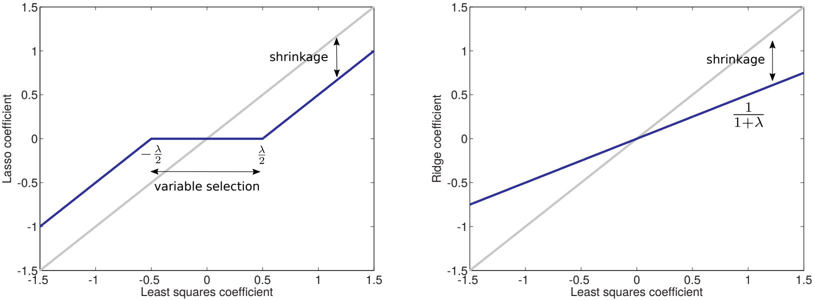
$$J^r(\beta) = ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{12}$$

- $\lambda \geq 0$ is the regularization (shrinkage) parameter.
- Penalty: sum of *absolute* values instead of sum of squares
- Difference seems minor but it results in a very different behavior: it enables *shrinkage* and *selection* of covariates.

Shrinkage and variable selection with the lasso

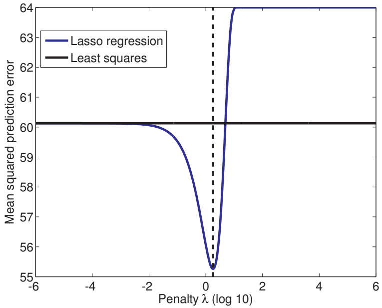
- The lasso generally lacks an analytical solution.
- Closed form solution when $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$

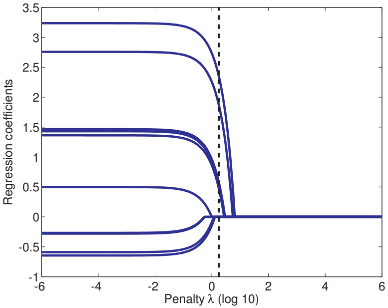
$$\hat{\beta}_j^L = \begin{cases} \hat{\beta}^o - \frac{\lambda}{2} & \text{if } \hat{\beta}^o \geq \frac{\lambda}{2} \\ 0 & \text{if } \hat{\beta}^o \in (-\frac{\lambda}{2},\ \frac{\lambda}{2}) \\ \hat{\beta}^o + \frac{\lambda}{2} & \text{if } \hat{\beta}^o \leq -\frac{\lambda}{2} \end{cases} \tag{13}$$



Back to the example

- Data were artificially generated with $\beta^* = (3, 2, 1, 0, \ldots, 0)^\top$
- The vector is sparse: only 3/10 nonzero terms
- Lasso regression combines shrinkage and variable selection.





Proof

- Assume $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$
- We want to show that the shrinkage and selection operator

$$\hat{\beta}_j^L = \begin{cases} \hat{\beta}^o - \frac{\lambda}{2} & \text{if } \hat{\beta}^o \geq \frac{\lambda}{2} \\ 0 & \text{if } \hat{\beta}^o \in (-\frac{\lambda}{2},\ \frac{\lambda}{2}) \\ \hat{\beta}^o + \frac{\lambda}{2} & \text{if } \hat{\beta}^o \leq -\frac{\lambda}{2} \end{cases} \tag{14}$$

minimizes $J^L(\beta)$,

$$J^L(\beta) = ||\boldsymbol{y} - \mathbf{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{15}$$

Proof

$$J^L(\beta) = ||\boldsymbol{y} - \mathbf{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{16}$$

$$= (\boldsymbol{y} - \mathbf{X}\beta)^\top (\boldsymbol{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{17}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{18}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{I}_p} \boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{19}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \underbrace{\mathbf{X}^\top \boldsymbol{y}}_{\hat{\beta}^o = \boldsymbol{r}} + \boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{20}$$

Proof

$$J^L(\beta) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{r} + \boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{21}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} - 2\sum_{j=1}^{p} \beta_j r_j + \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{22}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} + \sum_{j=1}^{p} \underbrace{\left(-2\beta_j r_j + \beta_j^2 + \lambda|\beta_j|\right)}_{f_j(\beta_j)} \tag{23}$$

$$= \text{constant} + \sum_{j=1}^{p} f_j(\beta_j) \tag{24}$$

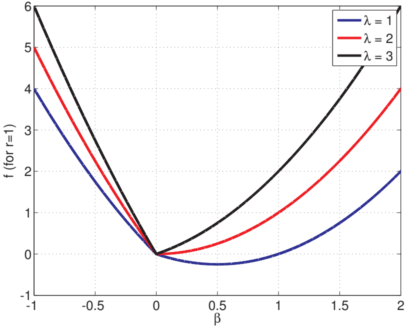
- For $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, the optimization problem decomposes into p independent problems.
- Minimizing each $f_j(\beta_j)$ separately will minimize $J^L(\beta)$.

Proof

- Drop the subscripts for a moment and consider a single f only.

$$f(\beta) = \beta^2 - 2r\beta + \lambda|\beta| \tag{25}$$

- Problem: derivative at zero not defined

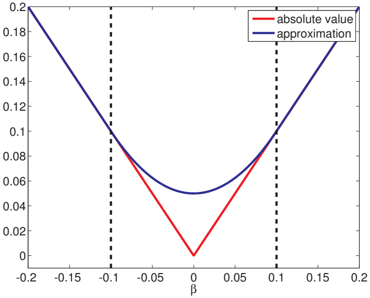


Proof

- Approach: Make a smooth approximation $|\beta| \approx h_\epsilon(\beta)$

$$h_\epsilon(\beta) = \begin{cases} \frac{\epsilon}{2} + \frac{1}{2\epsilon}\beta^2 & \text{if } \beta \in (-\epsilon,\ \epsilon) \\ |\beta| & \text{otherwise} \end{cases} \tag{26}$$

- Do all the work with $\epsilon > 0$ and, at the end, take the limit $\epsilon \to 0$.



Proof

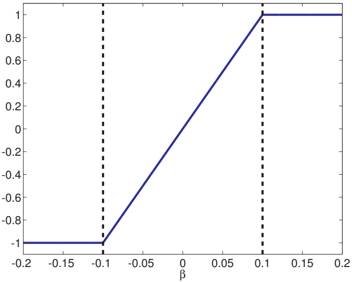
- Using $h_\epsilon(\beta)$ instead of $|\beta|$ gives

$$\tilde{f}(\beta) = \beta^2 - 2r\beta + \lambda h_\epsilon(\beta) \tag{27}$$

- The derivative of $\tilde{f}(\beta)$ is

$$\tilde{f}'(\beta) = 2\beta - 2r + \lambda h'_\epsilon(\beta) \tag{28}$$

$$h'_\epsilon(\beta) = \begin{cases} 1 & \text{if } \beta \geq \epsilon \\ \frac{\beta}{\epsilon} & \text{if } \beta \in (-\epsilon,\ \epsilon) \\ -1 & \text{if } \beta \leq -\epsilon \end{cases}$$



Proof

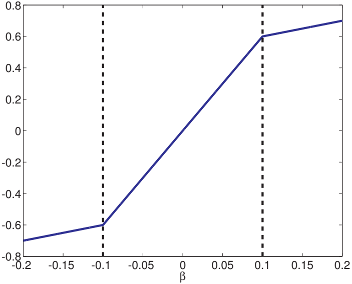
- Setting the derivative of $\tilde{f}'(\beta)$ to zero gives the condition

$$2\beta - 2r + \lambda h'_\epsilon(\beta) = 0 \quad (29)$$

$$\beta + \frac{\lambda}{2} h'_\epsilon(\beta) = r \quad (30)$$

- The left-hand side is a piecewise linear, monotonically increasing function $g_\epsilon(\beta)$: β is uniquely determined by r.

$$g_\epsilon(\beta) = \begin{cases} \beta + \frac{\lambda}{2} & \text{if } \beta \geq \epsilon \\ \beta(1 + \frac{\lambda}{2\epsilon}) & \text{if } \beta \in (-\epsilon,\ \epsilon) \\ \beta - \frac{\lambda}{2} & \text{if } \beta \leq -\epsilon \end{cases}$$



Proof

There are three cases

1. $r \geq \epsilon + \frac{\lambda}{2}$

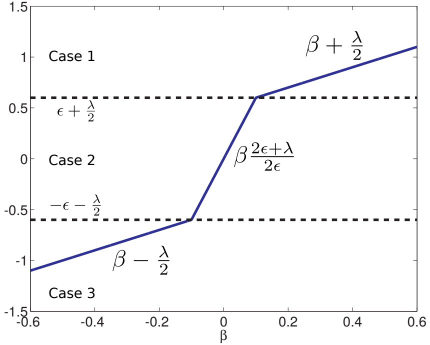
$$\beta + \frac{\lambda}{2} \stackrel{!}{=} r \;\Rightarrow\; \beta = r - \frac{\lambda}{2}$$

2. $r \in \left(-\epsilon - \frac{\lambda}{2},\ \epsilon + \frac{\lambda}{2}\right)$

$$\beta \frac{2\epsilon + \lambda}{2\epsilon} \stackrel{!}{=} r \;\Rightarrow\; \beta = \frac{2\epsilon r}{2\epsilon + \lambda}$$

3. $r \leq -\epsilon - \frac{\lambda}{2}$

$$\beta - \frac{\lambda}{2} \stackrel{!}{=} r \;\Rightarrow\; \beta = r + \frac{\lambda}{2}$$



Proof

There are three cases

1. $r \geq \epsilon + \frac{\lambda}{2}$

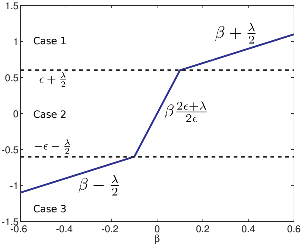
$$\beta + \frac{\lambda}{2} \stackrel{!}{=} r \;\Rightarrow\; \beta = r - \frac{\lambda}{2}$$

2. $r \in \left(-\epsilon - \frac{\lambda}{2},\, \epsilon + \frac{\lambda}{2}\right)$

$$\beta \frac{2\epsilon + \lambda}{2\epsilon} \stackrel{!}{=} r \;\Rightarrow\; \beta = \frac{2\epsilon r}{2\epsilon + \lambda}$$

3. $r \leq -\epsilon - \frac{\lambda}{2}$

$$\beta - \frac{\lambda}{2} \stackrel{!}{=} r \;\Rightarrow\; \beta = r + \frac{\lambda}{2}$$



Hence

$$\beta = \begin{cases} r - \frac{\lambda}{2} & \text{if } r \geq \epsilon + \frac{\lambda}{2} \\ \frac{2\epsilon}{2\epsilon+\lambda} r & \text{if } r \in \left(-\epsilon - \frac{\lambda}{2},\, \epsilon + \frac{\lambda}{2}\right) \\ r + \frac{\lambda}{2} & \text{if } r \leq -\epsilon - \frac{\lambda}{2} \end{cases}$$

Proof

- Taking the limit $\epsilon \to 0$ gives

$$\hat{\beta} = \begin{cases} r - \frac{\lambda}{2} & \text{if } r \geq \frac{\lambda}{2} \\ 0 & \text{if } r \in (-\frac{\lambda}{2}, \ \frac{\lambda}{2}) \\ r + \frac{\lambda}{2} & \text{if } r \leq -\frac{\lambda}{2} \end{cases} \tag{31}$$

- With the subscripts, and $r_j = \hat{\beta}_j^o$, we have

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^o - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^o \geq \frac{\lambda}{2} \\ 0 & \text{if } \hat{\beta}_j^o \in (-\frac{\lambda}{2}, \ \frac{\lambda}{2}) \\ \hat{\beta}_j^o + \frac{\lambda}{2} & \text{if } \hat{\beta}_j^o \leq -\frac{\lambda}{2} \end{cases} \tag{32}$$

which is the lasso solution $\hat{\beta}_j^L$.

Summary

Lasso $\equiv$ *Least Absolute Shrinkage and Selection Operator*

- Method to regularize linear regression (like ridge regression)
- Regularization / *shrinkage* can improve prediction accuracy.
- Method to perform covariate *selection* (unlike ridge regression)
- Covariate selection reduces the complexity of fitted models; makes them easier to interpret.
- Combination of shrinkage and selection is achieved by penalizing the *absolute* values of the regression coefficients.

Appendix

Constrained optimization point of view

Ridge regression:

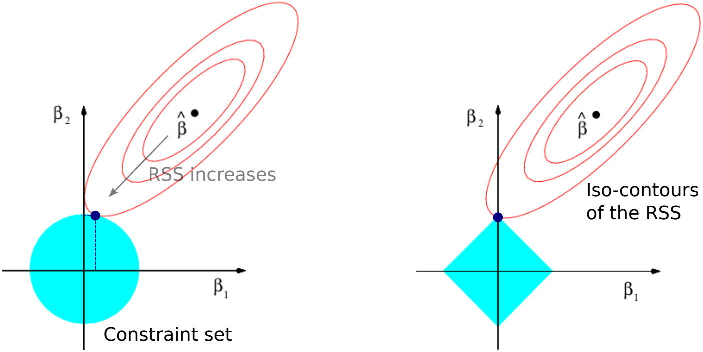
$$\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \mathbf{X}\beta||_2^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

Lasso regression:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \mathbf{X}\beta||_2^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$



(Based on figures from chapter 6 of *Introduction to Statistical Learning*)