# Information and Model Selection

Vahid Partovi Nia

Advanced Machine Learning: Lecture 03

February 25, 2018

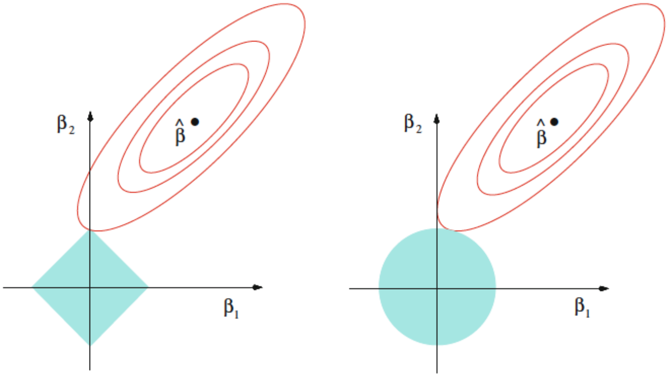
# Outline

1. Least Angle
2. Information

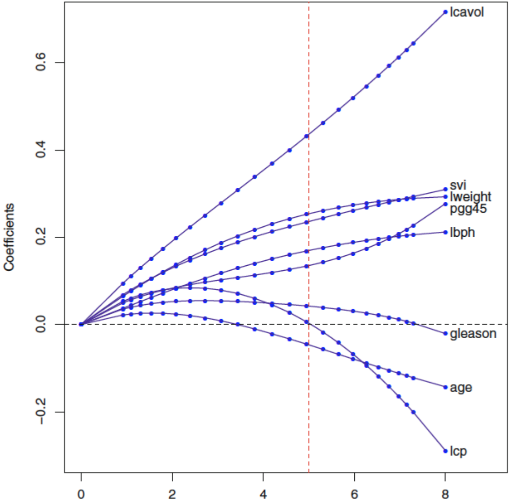
# Multiple Regression

1. initialize $\mathbf{r}_0 = \mathbf{1}$
2. $j = 1$
3. regress $\mathbf{x}_j$ on $(\mathbf{r}_0, \ldots, \mathbf{r}_{j-1})$
4. $\hat{\gamma}_{lj} = \frac{\mathbf{r}_l^\top \mathbf{x}_j}{\mathbf{r}_l^\top \mathbf{r}_l}$
5. orthogonalize $\mathbf{r}_j = \mathbf{x}_j - \sum_{k=1}^{j-1} \hat{\gamma}_{kj} \mathbf{r}_k$
6. $j = j + 1$ go to 3
7. $\hat{\beta}_p = \frac{\mathbf{y}^\top \mathbf{r}_p}{\mathbf{r}_p^\top \mathbf{r}_p}$

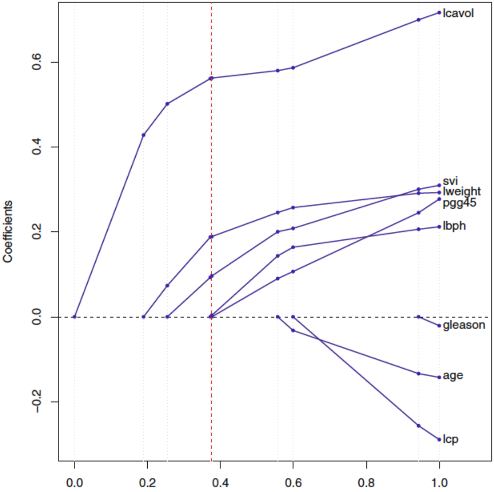
# Visual penalization



# Ridge coefficients



# Lasso coefficients



# Penalization and Selection

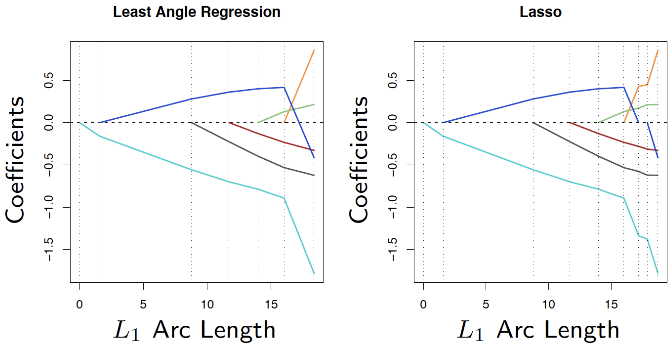
$$S(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})$$

- $P_\lambda(\boldsymbol{\beta})$ must be non-differentiable on the axes to select.
- Lasso has partially linear path. This helps to develop the path algorithm.
- Linear selection path appears while $S(\boldsymbol{\beta})$ is partially quadratic with non-differentiable penalization on axes.

1. standardize $\mathbf{x}_j$
2. set $\beta_j = 0, j = 1, \ldots, p,$
3. initialize $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$
4. Find the most correlated $\mathbf{x}_j$ with $\mathbf{r}\hat{j} = \mathrm{argmax}\ r^\top \mathbf{x}_j$
5. Move $\beta_j$ towards its least squares, $\beta_j = \delta \frac{\mathbf{r}^\top \mathbf{x}_j}{\mathbf{x}_j^\top \mathbf{x}_j}$
6. Update residual $\mathbf{r} = \mathbf{y} - \beta_j \mathbf{x}_j$ until $\mathbf{x}_k$ have more correlation.
7. Continue until $p$ predictors are in.

1. standardize $\mathbf{x}_j$
2. set $\beta_j = 0, j = 1, \ldots, p, k = 0, A_k = \emptyset$
3. initialize $\mathbf{r}_k = \mathbf{y} - \bar{\mathbf{y}}$
4. Add the most correlated predictor to $A_k$.
5. $\boldsymbol{\beta}_{A_k} = (\mathbf{X}_{A_k}^\top \mathbf{X}_{A_k})^{-1} \mathbf{X}_{A_k}^\top \mathbf{y}$
6. $\mathbf{r}_{A_k} = \mathbf{y} - \mathbf{X}_{A_k} \boldsymbol{\beta}_{A_k}$
7. $\delta_{A_k} = (\mathbf{X}_{A_k}^\top \mathbf{X}_{A_k})^{-1} \mathbf{X}_{A_k}^\top \mathbf{r}_{A_k}$
8. $\boldsymbol{\beta}_{A_k}(\delta) = \boldsymbol{\beta}_{A_k} + \delta \times \delta_{A_k}$, increase $\delta$ until another predictor (out of $A_k$) is more correlated with $\mathbf{y}$. Go to 4.

# LAR


Least Angle Regression
Lasso
Coefficients
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
$L_1$ Arc Length

# LAR Lasso

- If a nonzero coefficient hits zero, put variables out of $A_k$
- Recompute least squares.
- Go back to LAR algorithm.

# Other sparse estimators

Lasso:

$$\hat{\beta}^{\text{lasso}} \;=\; \operatorname{argmin}\, (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Dantzig:

$$\hat{\beta}^{\text{DS}} \;=\; \operatorname{argmin}\, ||\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_\infty + \lambda \sum_{j=1}^{p} |\beta_j|$$

# Importance sampling

$$\int_0^1 h(x)dx$$

$$\int_{-\infty}^{\infty} h(x)dx$$

# Likelihood

- Why do we need parametric models?
- Why do we use likelihood?
- Why maximum likelihood is good?
- What information means?
- How information is related to data?

Least Angle
Information

$\mathbb{KL}$ divergence between the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\begin{aligned} \mathbb{KL}(\theta_0, \theta) &= \int \log\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} f(x \mid \theta_0) \\ &= \mathbb{E}_{\theta_0}\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} \end{aligned}$$

Least Angle
Information

$\mathbb{KL}$ divergence between the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\begin{aligned}\mathbb{KL}(\theta_0, \theta) &= \int \log\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} f(x \mid \theta_0) \\ &= \mathbb{E}_{\theta_0}\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\}\end{aligned}$$

$$\mathbb{KL}(\theta_0, \theta) \neq \mathbb{KL}(\theta, \theta_0)$$

Cross entropy of the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\mathbb{H}(\theta, \theta_0) = \int \log f(x \mid \theta) f(x \mid \theta_0) dx$$

# KL and H

$$\mathbb{KL}(\theta_0, \theta) = \mathbb{H}(\theta_0, \theta_0) - \mathbb{H}(\theta, \theta_0)$$

# Properties

- $\mathbb{KL}(\theta_0, \theta) > 0$ iff $f(x \mid \theta_0) \neq f(x \mid \theta)$ on a set of $x$ with positive measure.
- $\mathbb{KL}(\theta_0, \theta) = 0$ iff $f(x \mid \theta_0) = f(x \mid \theta)$ almost everywhere.
- $\mathbb{KL}_n(\theta_0, \theta) = n\mathbb{KL}(\theta_0, \theta)$ for a set of i.i.d observations $(x_1, \ldots, x_n)$.
- $\frac{\partial \mathbb{H}(\theta, \theta_0)}{\partial \theta}\Big|_{\theta=\theta_0} = 0$
- $\frac{\partial^2 \mathbb{H}(\theta, \theta_0)}{\partial \theta \partial \theta^\top}\Big|_{\theta=\theta_0} = -J(\theta_0)$ where $J(.)$ is the observed information.

# More about entropy

Suppose $A = \{A_1, \ldots, A_k\}$ with probabilities $p_1, \ldots, p_k$.
Define $A'$ to be an $A$-similar event as
$A' = \{A_1, \ldots, A_k, A_{k+1}\}$ with probabilities
$p_1, \ldots, p_k, p_{k+1} = 0$.

- If two sets $A$ and $B$ are independent $\mathbb{H}(A \times B) = \mathbb{H}(A) + \mathbb{H}(B)$.
- $\mathbb{H}(A) = \mathbb{H}(A')$ .

The only function that satisfies the above two properties is $\mathbb{H}(A) = \lambda \sum_i p_i \log p_i$.
Why this result is important?

## More about entropy

| $A_1$ | $A_2$ |
|---|---|
| 0.1 | 0.9 |
| 0.49 | 0.51 |
| 0.69 0.325 | |