ST509 Computational Statistics

Lecture 5: LASSO and Its Computation

Seung Jun Shin

Department of Statistics
Korea University

E-mail: sjshin@korea.ac.kr

Stein's Paradox I

- Let $\mathbf{X} = (X_1, \cdots, X_p)^T \sim N_p(\boldsymbol{\theta}, \mathbf{I})$.
- The UMVUE and MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^0(\mathbf{X}) = \mathbf{X},$$

  which is the most obvious estimator that possess desirable properties.
- Nevertheless, Stein (1956) showed that $\hat{\boldsymbol{\theta}}^0(\mathbf{X})$ is inadmissible for squared error loss when $p \geq 3$, where

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sum_{i=1}^{p}(\hat{\theta}_i - \theta_i)^2$$

Stein's Paradox II

- James and Stein (1961) showed that

$$\hat{\boldsymbol{\theta}}^{JS}(\mathbf{X}) = \left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}$$

strictly dominates $\hat{\boldsymbol{\theta}}^{0}(\mathbf{X})$. (James-Stein Estimator)

proof The risk of JS estimator is

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}^{JS}, \boldsymbol{\theta}) &= E\left\{\left\|\mathbf{X} - \theta - \frac{(p-2)\mathbf{X}}{\|\mathbf{X}\|^2}\right\|^2\right\} \\ &= p - 2(p-2)\sum_{j=1}^{p} E\left\{\frac{X_j(X_j - \theta_j)}{\|\mathbf{X}\|^2}\right\} + (p-2)^2 E\left(\frac{1}{\|\mathbf{X}\|^2}\right) \end{aligned}$$

It turns out that

$$\sum_{j=1}^{p} E\left\{\frac{X_j(X_j - \theta_j)}{\|\mathbf{X}\|^2}\right\} = (p-2)E\left(\frac{1}{\|\mathbf{X}\|^2}\right)$$

Stein's Paradox III

since

$$
\begin{aligned}
& E\left\{\frac{X_1(X_1-\theta_1)}{\|\mathbf{X}\|^2}\right\} \\
=& \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{x_1}{\|\mathbf{x}\|^2} \times \frac{(x_1-\theta_1)}{(2\pi)^{p/2}} e^{-\|\mathbf{x}-\boldsymbol{\theta}\|^2/2} dx_1\cdots dx_p \\
=& \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\|\mathbf{x}\|^2-2x_1^2}{\|\mathbf{x}\|^4} \times \frac{1}{(2\pi)^{p/2}} e^{-\|\mathbf{x}-\boldsymbol{\theta}\|^2/2} dx_1\cdots dx_p \\
=& E\left(\frac{\|\mathbf{X}\|^2-2X_1^2}{\|\mathbf{X}\|^4}\right).
\end{aligned}
$$

Finally, we have

$$
R(\hat{\boldsymbol{\theta}}^{JS},\boldsymbol{\theta}) = p-(p-2)E\left(\frac{1}{\|\mathbf{X}\|^2}\right) < p. \qquad \blacksquare
$$

Stein's Paradox IV

- JS estimator shrinks each component of X towards the origin, and thus the biggest improvement comes when $\|\boldsymbol{\theta}\|$ is close to zero.
- Normality assumption is not critical, and similar results can be shown for a wide class of distributions.
- Stein phenomenon holds only when our interest is in the simultaneous estimation of all components in θ.

Ridge Regression I

- Linear Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

  where $\mathbf{e}$ be random errors with $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$.

- OLS estimator solves

$$\hat{\boldsymbol{\beta}}_{\text{ols}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

  and satisfies

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{y}.$$

- If $\mathbf{X}$ is of full rank, we have

$$\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Ridge Regression II

- Ridge Regression estimator is

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

  for a given δ.

- RR originally proposed for mitigate the collinearity in $\mathbf{X}$.
- It is not difficult to show that (Unconstraint Form):

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{2n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$$

  where $\|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}$.

- Equivalently, we can rewrite (Constraint Form):

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad \text{s.t. } \|\boldsymbol{\beta}\|_2^2 \leq t.$$

Ridge Regression III

- Prediction Error for $y_0 = \mu_0 + \epsilon_0 = \mathbf{x}_0^T \boldsymbol{\beta} + \epsilon_0$:

$$E\{(y_0 - \hat{y}_0)^2\} = \underbrace{\sigma^2}_{\text{Irreducible}} + \underbrace{E(\mu_0 - \hat{y}_0)^2}_{\text{Model Error}}$$

  where $\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$.

- Model error can be decomposed as:

$$\begin{aligned} E(\mu_0 - \hat{y}_0)^2 &= E\left\{\mu_0 - E(\hat{y}_0) + E(\hat{y}_0) - \hat{y}_0\right\}^2 \\ &= \underbrace{\{\mu_0 - E(\hat{y}_0)\}^2}_{\text{Bias}^2} + \underbrace{\text{Var}(\hat{y}_0)}_{\text{Variance}} \end{aligned}$$

Ridge Regression IV

- Principle: The larger (smaller) model space, the larger (smaller) variance and the smaller (larger) bias of the fitted model.
- RR restricts the parameter space from $\mathbb{R}^p$ to $\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_2^2 \leq t\}$.
- Compared to $\hat{\boldsymbol{\beta}}_{\text{ols}}$ known to be BLUE, $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ is biased but tends to have smaller variance.
- Thus $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ with a carefully selected λ (or t) can beat $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

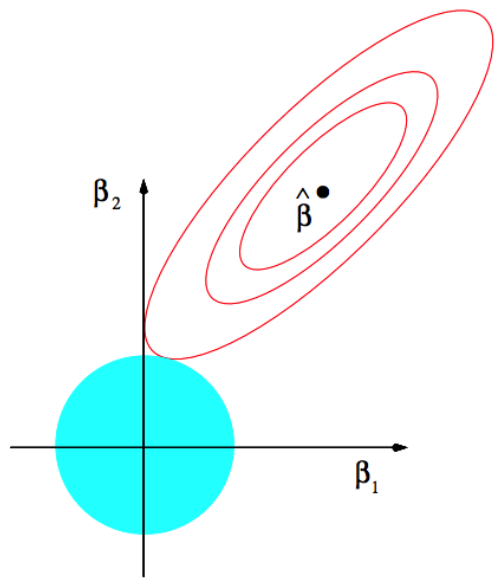


Figure: Geometry of Ridge Regression (from ELS)

LASSO I

- LASSO proposed by Tibshirani (1996) solves (unconstraint form):

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$$

  where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

- Its constraint form is given by

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad \text{s.t. } \|\boldsymbol{\beta}\|_1 \leq t$$

  for some t.

LASSO II

- LASSO produces a sparse solution and thus achieves variable selection and estimation simultaneously.
- LASSO can beat OLS estimator, just like RR.

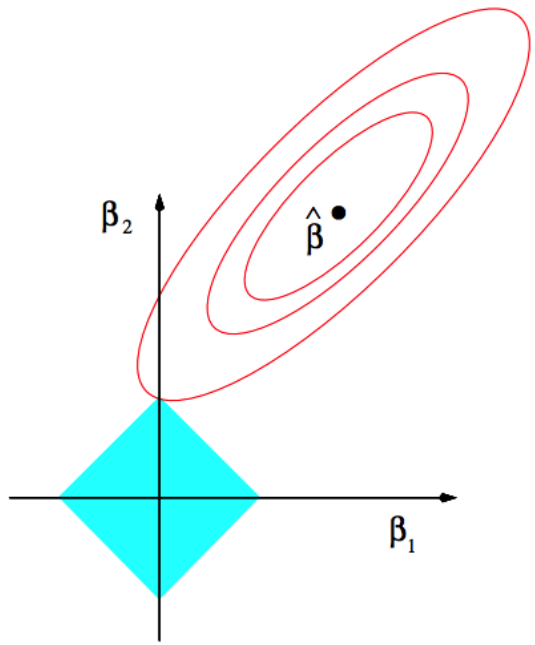


Figure: Geometry of LASSO (from ELS)

LASSO III

- Let z_i be the standardized s.t.

$$\sum_{i=1}^{n} z_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} z_i^2 = n$$

- OLS estimator is

$$\hat{\beta}_{\text{ols}} = \frac{1}{n} \sum y_i z_i$$

- RR estimator is

$$\hat{\beta}_{\text{ridge}} = \hat{\beta}_{\text{ols}} / (1 + \lambda)$$

- LASSO estimator is

$$\hat{\beta}_{\text{lasso}} = \begin{cases} \hat{\beta}_{\text{ols}} - \lambda & \text{if } \hat{\beta}_{\text{ols}} > \lambda \\ 0 & \text{if } \mid \hat{\beta}_{\text{ols}} \mid \leq \lambda \\ \hat{\beta}_{\text{ols}} + \lambda & \text{if } \hat{\beta}_{\text{ols}} < -\lambda \end{cases}$$

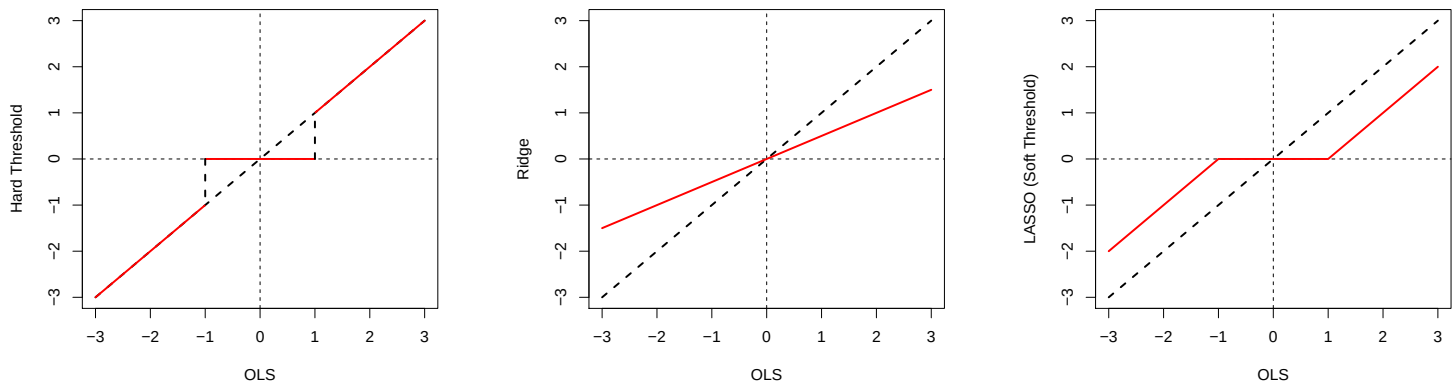
LASSO IV

- Define the soft-thresholding operator, $S_\lambda(x)$ as

$$S_\lambda(x) = \text{sign}(x)\left(|x| - \lambda\right)_+ .$$

- Then we have

$$\hat{\beta}_{\text{lasso}} = S_\lambda(\hat{\beta}_{\text{ols}})$$



(a) Hard Thresh. (b) Ridge Regression (c) Lasso (Soft Thresh.)

Figure: Comparison to OLS: one-dimensional orthogonal regression.

Computation I

- Our goal is to solve

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

  or

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2, \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| < t$$

- This is standard form of Quadratic Programing (QP).

Computation II

- Wu & Lange (2008) proposed the coordinate decent algorithm to solve LASSO.

1. Initialize $\boldsymbol{\beta}^{(0)}$.
2. At the t^{th} iteration, we update $\beta_j, j \in \{1, \cdots, p\}$ in a cyclic way;

$$\beta_j^{(t+1)} = \underset{\beta_j}{\operatorname{argmin}} f(\beta_1^{(t+1)}, \cdots, \beta_{j-1}^{(t+1)}, \beta_j, \beta_{j+1}^{(t)}, \cdots, \beta_p^{(t)})$$

3. Repeat the above until convergence.

Algorithm 1: (Cyclic) Coordinate Decent Algorithm to solve $\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$.

Computation III

- Stochastic CD algorithm randomly chooses the coordinate j to be updated.
- CD algorithm is thus particularly useful when one-dimensional problem has a closed form of solution.
- CD can be viewed as a version of the gradient decent algorithm.

Computation IV

- For LASSO problem, we have the following updating equation for β_j

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n}(\mathbf{r}_{-j} - \beta_j \mathbf{x}_j)^T(\mathbf{r}_{-j} - \beta_j \mathbf{x}_j) + \lambda|\beta_j|$$

where

$$\mathbf{r}_{-j} = \mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}, \qquad \text{(partial residual)}$$

with $\hat{\boldsymbol{\beta}}$ is the most recently updated value of $\boldsymbol{\beta}$.

- The updating equation is

$$\hat{\beta}_j = S_\lambda\left(\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_{-j}\right) = S_\lambda\left(\hat{\beta}_j + \frac{1}{n}\mathbf{x}_j^T\mathbf{r}\right)$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ (full residual).

Computation V

1. WLOG, $\mathbf{x}_i$ are marginally standardized and y_i are centered such that

$$\sum_{i=1}^{n} x_{ij} = 0, \sum_{i=1}^{n} x_{ij}^2 = n, \text{ and } \sum_{i=1}^{n} y_i = 0$$

   You may need to compute $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, $s_j = \sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2/n}$, and $\bar{y} = \sum_{i=1}^{n} y_i/n$.

2. Initialize $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ and compute $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

3. For $j = 1, \cdots, p$,

   3.1 Update coefficients:

$$\hat{\beta}_j^{(t+1)} = S_{\lambda/s_j}\left(\hat{\beta}_j^{(t)} + \frac{1}{n}\mathbf{x}_j^T\mathbf{r}\right).$$

   3.2 Update residuals:

$$\mathbf{r} = \mathbf{r} - \left(\beta_j^{(t+1)} - \beta_j^{(t)}\right)\mathbf{x}_j.$$

4. Repeat Step 2.1–2 until convergence.

5. Transform back to the original scale: $\hat{\beta}_0 \leftarrow \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j/s_j$ and $\hat{\beta}_j \leftarrow \hat{\beta}_j/s_j$.

Computation VI

- Convergence of CD algorithm is guaranteed by that the objective function is coutinuously differentiable and strictly convex in each coordinate.
- Suppose that the objective function $f(\boldsymbol{\beta})$ is

$$f(\beta_1, \cdots, \beta_p) = g(\beta_1, \cdots, \beta_p) + \sum_{j=1}^{p} h_j(\beta_j)$$

  where
  - g is differentiable and convex;
  - h_j is convex (but not necessarily differentiable).
- Then CD converges to the global minimum $\boldsymbol{\beta}^* = \operatorname{argmin} f(\boldsymbol{\beta})$.

Tuning Parameter Selection I

- In LASSO, λ (tuning parameter) plays a crucial role for performance.
- Too large λ under-fits the model and too small λ over-fits the model.
- This is generally true for other penalized methods beyond LASSO.
- Grid search via Cross-validation (CV) is popular.
- However, CV is often too computationally heavy, due to the numerical optimization should be repeatedly carried out for many different λs in the grid.
- More importantly, the choice of grid is often very subjective for both range and coarseness.

Tuning Parameter Selection II

- To reduce computational burden, Friedman et al. (2007) proposed pathwise coordinate optimization.
- The idea is to apply a CD procedure for each value of λ , varying λ along the path.
- Each solution is used as a warm start for the next problem.

LARS I

- Efron et al (2004) proposed.
- Possible to compute entire path of $\boldsymbol{\beta}_{LASSO}(\lambda)$ as a function of λ.
- Forward addition sequentially builds a model by adding one variable at a time; identifies variables ($\in \mathcal{S}$), then computes OLS for all variables $\mathcal{S}$.
- LARS shares a similar spirit to FA, but a sort of democratic version.

LARS II

- Without loss of generality, we assume

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1 \qquad \text{for } j = 1, \cdots, p.$$

- Let $\hat{\boldsymbol{\mu}}_{\mathcal{S}}$ denote a current model estimate, then the correlations are

$$\hat{\mathbf{c}}(\hat{\boldsymbol{\mu}}_{\mathcal{S}}) = \mathbf{X}_{\mathcal{S}^c}^T \mathbf{r}(\hat{\boldsymbol{\mu}}_{\mathcal{S}}) = \mathbf{X}_{\mathcal{S}^c}^T (\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{S}})$$

- Forward Stagewise (FS) updates $\hat{\boldsymbol{\mu}}$:

$$\hat{\boldsymbol{\mu}}_{\mathcal{S} \cup \hat{j}} \to \hat{\boldsymbol{\mu}}_{\mathcal{S}} + \epsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}}$$

  where $\hat{j} = \text{argmax}_{j \in \mathcal{S}^c} |\hat{c}_j|$ for a given $\epsilon > 0$.

- Classical Forward Addition updates

$$\hat{\boldsymbol{\mu}}_{\mathcal{S} \cup \hat{j}} \to \hat{\boldsymbol{\mu}}_{\mathcal{S}} + \hat{c}_{\hat{j}} \cdot \mathbf{x}_{\hat{j}}$$

LARS III

- LARS algorithm:

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}\cup\hat{j}} \to \hat{\boldsymbol{\mu}}_{\mathcal{S}} + \hat{\gamma}_{\hat{j}} \cdot \mathbf{u}_{\mathcal{S}}$$

  where

  - $\mathbf{u}_{\mathcal{S}}$: an equiangular vector for $\mathbf{X}_{\mathcal{S}}$, i.e., the angles between $\mathbf{u}_{\mathcal{S}}$ and $\mathbf{x}_j, j \in \mathcal{S}$ are identical.
  - $\hat{\gamma}_{\hat{j}}$: the smallest step (> 0) such that the new index $\hat{j}$ joins the active set, $\mathcal{S} \to \mathcal{S} \cup \hat{j}$.

LARS IV

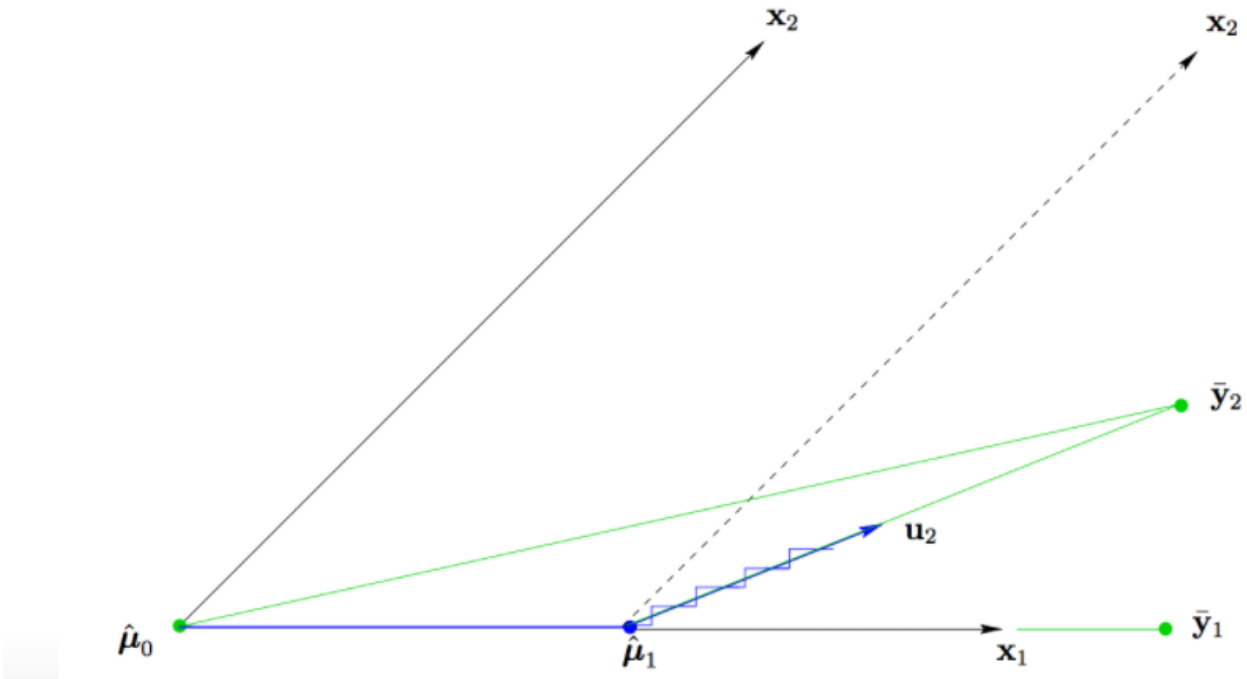


Figure: LARS on $(\mathbf{x}_1, \mathbf{x}_2)$;$\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into span$(\mathbf{x}_1, \mathbf{x}_2)$. Beginning at $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$, the residual vector $\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_0$ has greater correlation with $\mathbf{x}_1$ than $\mathbf{x}_2$; the next LARS estimate is $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}_1 \mathbf{x}_1$, where $\hat{\gamma}_1$ is chosen such that $\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$; then $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_2 \mathbf{u}_2$, where $\mathbf{u}_2$ is the unit bisector; $\hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$ in this case with two predictors.

LARS V

- LARS algorithm:
  1. Marginally transform the data with $\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} x_i^2 = 1$ and initialize $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ (or $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$).
  2. Let $\mathcal{S}$ be an index set of variables for the current step.
  3. Compute the equiangular unit vector $\mathbf{u}_{\mathcal{S}}$:

$$\mathbf{u}_{\mathcal{S}} = \tilde{\mathbf{X}}_{\mathcal{S}} \omega_{\mathcal{S}}, \qquad \omega_{\mathcal{S}} = A_{\mathcal{S}} G_{\mathcal{S}}^{-1} \mathbb{1}_{\mathcal{S}} \tag{2}$$

where

$$\begin{aligned}
\tilde{\mathbf{X}}_{\mathcal{S}} &= \{s_j \mathbf{x}_j ; j \in \mathcal{S}\} \text{ with some } s_j \in \{-1, 1\} \\
G_{\mathcal{S}} &= \tilde{\mathbf{X}}_{\mathcal{S}}^T \tilde{\mathbf{X}}_{\mathcal{S}} \\
A_{\mathcal{S}} &= (\mathbb{1}_{\mathcal{S}}^T G_{\mathcal{S}}^{-1} \mathbb{1}_{\mathcal{S}})^{-1/2}
\end{aligned}$$

LARS VI

4. Compute current correlation:

$$\hat{\mathbf{c}} = (\hat{c}_1, \cdots, \hat{c}_d)^T = \mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{S}})$$

and

$$\hat{C} = \max_j\{|\hat{c}_j|\} \quad \text{and} \quad \mathcal{S} = \{j : |\hat{c}_j| = \hat{C}\}$$

5. The LARS updates

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}+} = \hat{\boldsymbol{\mu}}_{\mathcal{S}} + \hat{\gamma}\mathbf{u}_{\mathcal{S}}$$

where $\mathbf{u}_{\mathcal{S}}$ is defined in (2) with $s_j = \text{sign}(\hat{c}_j)$ for $j \in \mathcal{S}$. Equivalently we can update $\boldsymbol{\beta}$ by (2)

$$\hat{\beta}_j^+ = \hat{\beta}_j + \hat{\gamma} \times s_j \times \omega_j, \quad \text{for } j \in \mathcal{S}$$

where ω_j is the jth element of $\omega_{\mathcal{S}}$.

LARS VII

6. The step size $\hat{\gamma}$ is chosen as

$$\hat{\gamma} = \min_{j \in \mathcal{S}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_{\mathcal{S}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathcal{S}} + a_j} \right\} \tag{3}$$

where $a_j = \mathbf{x}_j^T \mathbf{u}_{\mathcal{S}}$ where $j \in \mathcal{S}^c$.

- (Interpretation of (3)) Define for $\gamma > 0$

$$\boldsymbol{\mu}(\gamma) = \hat{\boldsymbol{\mu}}_{\mathcal{S}} + \gamma \mathbf{u}_{\mathcal{S}}$$

The current correlation between $\mathbf{y} - \mu(\gamma)$ and $\mathbf{x}_j, j \in \mathcal{S}^c$ is

$$c_j(\gamma) = \mathbf{x}_j^T(\mathbf{y} - \boldsymbol{\mu}(\gamma)) = \hat{c}_j - \gamma a_j, \qquad j \in \mathcal{S}^c. \tag{4}$$

We must have for $j \in \mathcal{S}$

$$\mathbf{c}_{\mathcal{S}}(\gamma) = \{c_j(\gamma); j \in \mathcal{S}\} = \mathbf{X}_{\mathcal{S}}^T(\mathbf{y} - \hat{\boldsymbol{\mu}} - \gamma \mathbf{u}_{\mathcal{S}}) = \mathbf{X}_{\mathcal{S}}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}) - \gamma \tilde{\mathbf{X}}_{\mathcal{S}}^T \mathbf{u}_{\mathcal{S}}$$

which yields

$$|c_j(\gamma)| = \hat{C} - \gamma A_{\mathcal{S}}, \quad \text{for } j \in \mathcal{S}. \tag{5}$$

Equating (4) and (5) yields (3).

7. Update $\mathcal{S} \to \mathcal{S} \cup \hat{j}$ where $\hat{j}$ is the index in $\mathcal{S}^c$ associated with $\hat{\gamma}$ in (3).

LARS VIII

- (Termination) In the last d step, (3) cannot be applied to get $\hat{\gamma}$, since $\mathcal{S}^c = \phi$, but we know that it will be OLS. After some calculation, we have $\hat{\gamma} = \hat{C}/A_{\mathcal{S}}$ in the final step.
- LARS takes only d steps and does NOT involve any numerical optimization.

Modification of LARS for LASSO I

- LARS and LASSO solution paths are nearly identical, but not exactly.

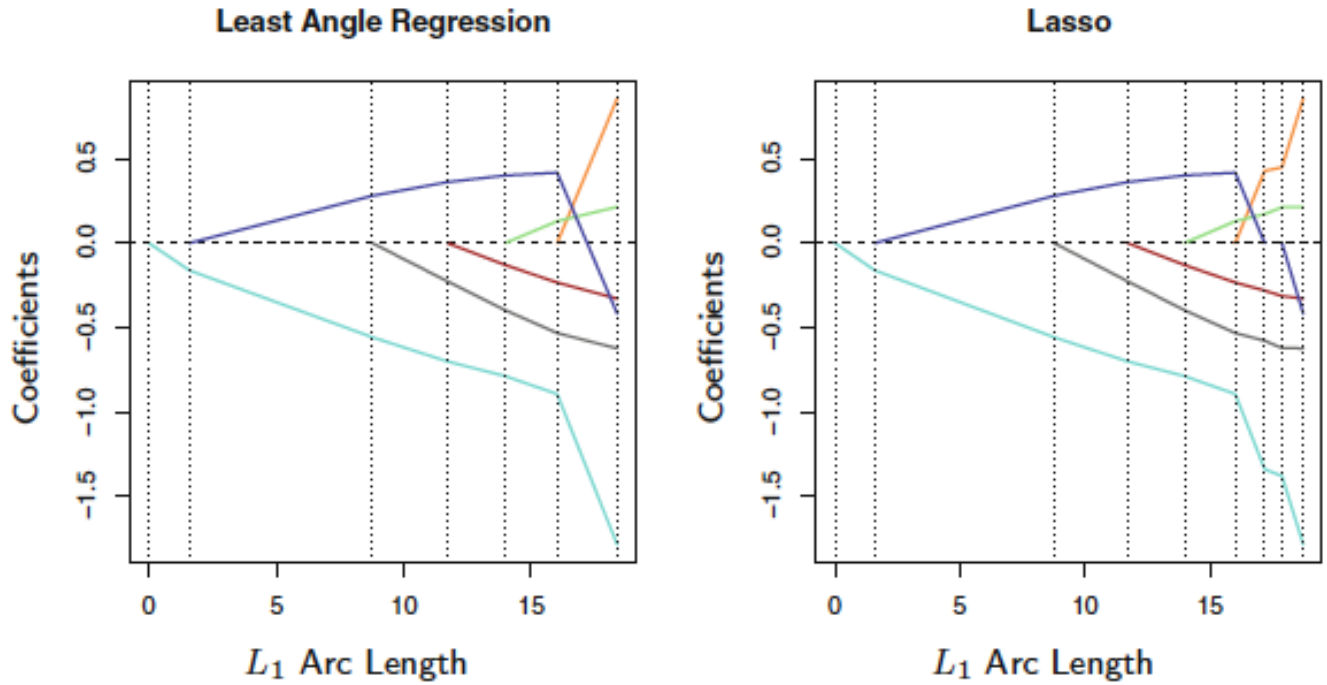


Figure: LARS vs LASSO: LASSO paths involves more steps due to additional deleting steps.

Modification of LARS for LASSO II

- In LARS, we have for all $j \in \mathcal{S}$,

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \gamma \times \text{sign}\{c_j\} \qquad \text{for some common } \gamma,$$

  whereas we have for all $j \in \mathcal{S}^c$

$$|\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})| \leq \gamma$$

- In LASSO, we have for all $j \in \mathcal{B} = \{j : \hat{\beta}_j \neq 0\}$,

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \lambda \times \text{sign}\{\beta_j\}$$

  and for $j \in \mathcal{B}^c$,

$$|\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})| \leq \lambda.$$

Modification of LARS for LASSO III

- Now you can see both similarity and difference between LARS and LASSO.
- LARS will produce different values if the sign of β_j is changed while the sign of correlation is not.
- LASSO modification: If a non-zero coefficient, say β_j hits zero, then drop it from $\mathcal{S} = \mathcal{S} \backslash j$ and recalculate direction $\mathbf{u}_{\mathcal{S}}$.

Degrees of Freedom of LASSO

- For $y_i|\mathbf{x}_i \overset{ind}{\sim} (f(\mathbf{x}_i), \sigma^2)$,

$$df(\hat{f}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(\hat{f}(\mathbf{x}_i), y_i)$$

- Under the linear model, we have

$$df(\hat{f}) = \text{tr}\left[\frac{1}{\sigma^2}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}(\mathbf{y}, \mathbf{y})\right] = \text{tr}(\mathbf{H}) = p$$

- Degrees of freedom of ridge regression is smaller than p even with p predictor.
- Degrees of freedom of LASSO:

$$df(\hat{f}(\lambda)) = |\mathcal{S}(\lambda)| = \#\text{ of nonzero coefficients.}$$

- Compare BIC and AIC at the knot points of the paths only.

Reference

- Friedman, Hastie, & Tibshirani (2009) The Elements of Statistical Learning; 2nd edition, Springer.
- Tibshirani, Wainwright, & Hastie (2015) Statistical Learning with Sparsity: the lasso and generalizations Chapman and Hall/CRC.
- Samworth (2008) Stein's Paradox. Eureka, 62, 38-41.
- Tibshirani (1996) Regression shrinkage and selection via the lasso JRSSb, 58(1), 267-288.
- Wu & Lange (2008) Coordinate descent algorithms for lasso penalized regression AOAS, 2(1), 224-244.
- Efron, Hastie, Johnstone, & Tibshirani (2004) Least angle regression (with Discussions) AOS, 32(2), 407-499
- Friedman, Hastie, Hfling, & Tibshirani (2007) Pathwise coordinate optimization AOAS, 1(2), 302-332.
- (2012) Zhou, Hastie, & Tibshirani (2007) On the "degrees of freedom" of the lasso AOS, 35(5), 2173-2192.