

Variable Selection

version: 2020-02-06 · 07:44:35

Motivation

- ▶ **Linear Regression Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$
 - ▶ n observations $\mathbf{y} = (y_1, \dots, y_n)$
 - ▶ p covariates $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$
- ▶ **Maximum Likelihood:** (assume $\sigma = 1$)
 - ▶ *Likelihood:* $\ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$.
 - ▶ For $p \leq n$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$. $\mathbf{X}'\mathbf{X}$ is almost surely invertible when $p \leq n$.

Motivation

► **Linear Regression Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$

► n observations $\mathbf{y} = (y_1, \dots, y_n)$

► p covariates $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$

► **Maximum Likelihood:** (assume $\sigma = 1$)

► *Likelihood:* $\ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$.

► For $p \leq n$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$. $\mathbf{X}'\mathbf{X}$ is almost surely invertible when $p \leq n$.

► For $p > n$: $\mathbf{X}'\mathbf{X}$ is singular \implies infinitely many likelihood-maximizing values:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\alpha}} + \text{span}\{\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_{p-n}\} \implies \sup \|\hat{\boldsymbol{\beta}}\| = \infty.$$

Motivation

▶ **Linear Regression Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$

▶ n observations $\mathbf{y} = (y_1, \dots, y_n)$

▶ p covariates $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$

▶ **Maximum Likelihood:** (assume $\sigma = 1$)

▶ *Likelihood:* $\ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$.

▶ For $p > n$: $\mathbf{X}'\mathbf{X}$ is singular \implies infinitely many likelihood-maximizing values:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\alpha}} + \text{span}\{\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_{p-n}\} \implies \sup \|\hat{\boldsymbol{\beta}}\| = \infty.$$

▶ **Application of $p > n$:**

▶ y_i : level of cancer-associated antigen in subject i .

▶ x_{ij} : expression level for gene j .

▶ Typical data: $n \sim 100 - 1\text{K}$, $p \sim 1\text{M}$.

Penalized Likelihood

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$

► **Unconstrained MLE:**

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$$

► **Penalized Likelihood:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) < t,$$

for some **penalty function** $\rho(\boldsymbol{\beta}) \geq 0.$

Penalized Likelihood

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$.

► **Unconstrained MLE:**

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$$

► **Penalized Likelihood:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) < t,$$

for some **penalty function** $\rho(\boldsymbol{\beta}) \geq 0$.

► For $p > n$: Typically $\rho(\boldsymbol{\beta}) < t \implies \|\boldsymbol{\beta}\| < C \implies \tilde{\boldsymbol{\beta}}$ is *likely* to be unique.

Penalized Likelihood

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$.

► **Unconstrained MLE:**

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$$

► **Penalized Likelihood:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) < t,$$

for some **penalty function** $\rho(\boldsymbol{\beta}) \geq 0$.

► For $p < n$: $\text{Bias}(\hat{\boldsymbol{\beta}}) = E[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}_{\text{true}} = 0$, but $\text{Bias}(\tilde{\boldsymbol{\beta}}) \neq 0$.

However, for $p \approx n$ $\text{var}(\tilde{\boldsymbol{\beta}}) \ll \text{var}(\hat{\boldsymbol{\beta}})$, such that PL can have smaller **mean squared error (MSE)**:

$$\text{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\text{true}}) \stackrel{\text{def}}{=} \sum_{i=1}^p E [(\hat{\theta}_i - \theta_{i,\text{true}})^2] = \sum_{i=1}^p \text{Bias}(\hat{\theta}_i)^2 + \text{var}(\hat{\theta}_i).$$

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$

► **L_2 -Constraint:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq t.$$

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2.$

► **L_2 -Constraint:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq t.$$

► **Standardization:** Preprocess \mathbf{X} and \mathbf{y} such that

► Columns $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})$ such that $\text{mean}(\mathbf{X}_j) = 0$, $\text{sd}(\mathbf{X}_j) = 1$.

► No intercept in model, i.e., $\text{mean}(\mathbf{y}) = 0$.

This is why the constraint has equally weighted β_j .

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$.

► **Penalized Likelihood:** Let $S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2$ and $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.

Constrained minimization

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) \leq t.$$

► **Unconstrained Formulation:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) + \lambda \cdot \rho(\boldsymbol{\beta}). \quad (1)$$

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$.

► **Penalized Likelihood:** Let $S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2$ and $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.

Constrained minimization

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) \leq t.$$

► **Unconstrained Formulation:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) + \lambda \cdot \rho(\boldsymbol{\beta}).$$

Proof: Let $\tilde{\boldsymbol{\beta}}$ be the solution to (1) and $t = \rho(\tilde{\boldsymbol{\beta}})$.

Then for $\boldsymbol{\beta}$ subject to $\rho(\boldsymbol{\beta}) \leq t = \rho(\tilde{\boldsymbol{\beta}})$,

$$\begin{aligned} S(\tilde{\boldsymbol{\beta}}) + \lambda \cdot \rho(\tilde{\boldsymbol{\beta}}) &\leq S(\boldsymbol{\beta}) + \lambda \cdot \rho(\boldsymbol{\beta}) \leq S(\boldsymbol{\beta}) + \lambda \cdot \rho(\tilde{\boldsymbol{\beta}}) \\ &\implies S(\tilde{\boldsymbol{\beta}}) \leq S(\boldsymbol{\beta}). \end{aligned}$$

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$.

► **Parameter Estimation:**

► *Unconstrained formulation:*

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

► *Solution:* PL is quadratic in $\boldsymbol{\beta}$, so use complete-the-squares to obtain

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda^2 \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y}.$$

Ridge Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$.

► **Parameter Estimation:**

► *Unconstrained formulation:*

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

► *Solution:* PL is quadratic in $\boldsymbol{\beta}$, so use complete-the-squares to obtain

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda^2 \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y}.$$

► **Question:** How to pick λ ?

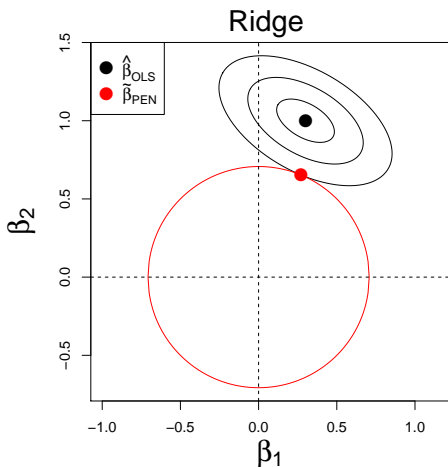
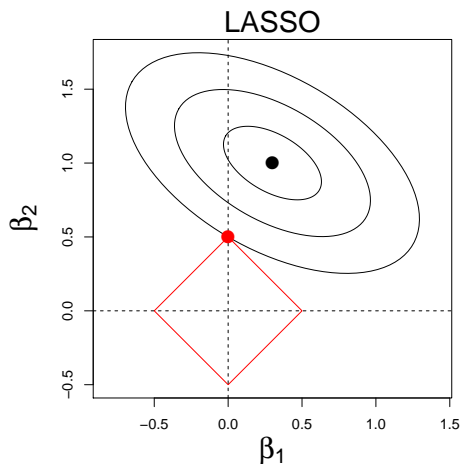
Lasso Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$.

► **L_1 Constraint:**

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq t \\ &= \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|.\end{aligned}$$

Lasso Regression



Advantage of Lasso over Ridge: [Variable selection](#), i.e., some of the $\tilde{\beta}_j$ in Lasso can equal 0. The black contours correspond to the shape of $S(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2$. The corners of the constraint region allow Lasso to perform variable selection.

Elastic Net Regression

- ▶ **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$.
- ▶ **Penalized Likelihood:** $\tilde{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) - \lambda \rho(\boldsymbol{\beta})$.
 - ▶ *Ridge Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.
 - ▶ *Lasso Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$.

Elastic Net Regression

- ▶ **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$.
- ▶ **Penalized Likelihood:** $\tilde{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) - \lambda \rho(\boldsymbol{\beta})$.
 - ▶ *Ridge Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.
 - ▶ *Lasso Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$.
- ▶ **Advantage of Lasso over Ridge:** **Variable selection**, i.e., some of the $\tilde{\beta}_j$ in Lasso can equal 0.

Elastic Net Regression

- ▶ **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$.
- ▶ **Penalized Likelihood:** $\tilde{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) - \lambda \rho(\boldsymbol{\beta})$.
 - ▶ *Ridge Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.
 - ▶ *Lasso Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$.
- ▶ **Advantage of Lasso over Ridge:** **Variable selection**, i.e., some of the $\tilde{\beta}_j$ in Lasso can equal 0.
- ▶ **Advantage of Ridge over Lasso:** Better performance when covariates are **highly correlated**.

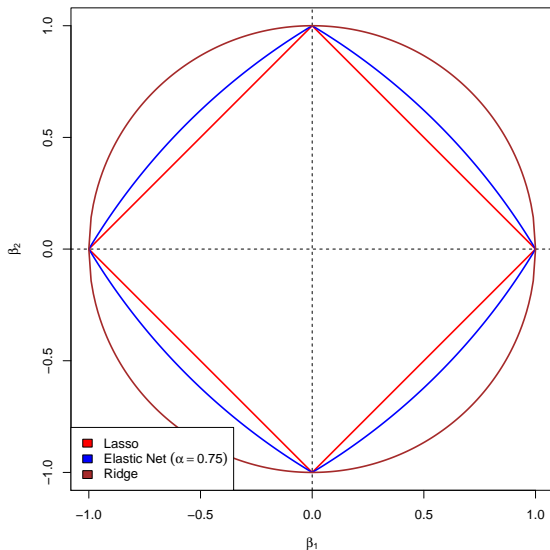
Elastic Net Regression

- ▶ **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$.
- ▶ **Penalized Likelihood:** $\tilde{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) - \lambda \rho(\boldsymbol{\beta})$.
 - ▶ *Ridge Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.
 - ▶ *Lasso Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$.
- ▶ **Elastic Net Regression:** Compromise between L_1 and L_2 constraints:

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \cdot \rho_{\alpha}(\boldsymbol{\beta}),$$

where $\rho_{\alpha}(\boldsymbol{\beta}) = \sum_{j=1}^p (1 - \alpha) \beta_j^2 + \alpha |\beta_j|$ for $\alpha \in [0, 1]$.

Elastic Net Regression



Constraint shapes for different penalty functions.

Elastic Net Regression

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$.

► **Parameter Estimation:**

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \underbrace{\sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}_{S(\boldsymbol{\beta})} + \lambda \cdot \underbrace{\sum_{j=1}^p (1 - \alpha) \beta_j^2 + \alpha |\beta_j|}_{\rho_{\alpha}(\boldsymbol{\beta})}.$$

► Lots of methods of solution, but the simplest and most effective for $p \gg n$ (and fixed λ, α) is **Coordinate Descent**:

- Minimize $\Omega(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + \lambda \rho_{\alpha}(\boldsymbol{\beta})$ one β_j at a time holding the others fixed.
- Continue cycling through β_j 's until relative tolerance is reached.

Elastic Net Regression

Coordinate Descent

- ▶ Minimize $\Omega(\beta_1, \beta^*)$ as a function of β_1 with $\beta^* = (\beta_2, \dots, \beta_p)$ fixed:

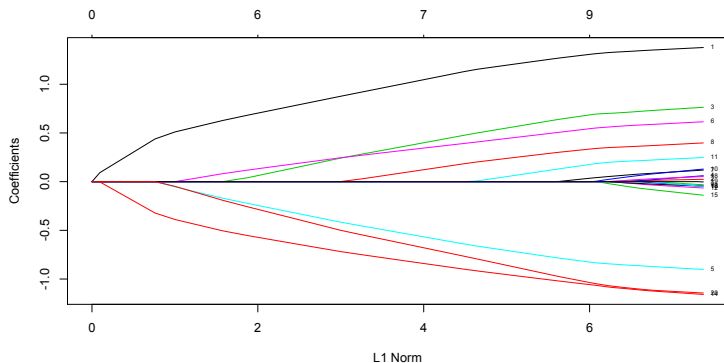
$$\begin{aligned}\Omega(\beta_1, \beta^*) &= \sum_{i=1}^n (y_i - x_{1i}\beta_1 - \mathbf{x}_i^* \beta^*)^2 + \lambda\alpha|\beta_1| + \lambda(1-\alpha)\beta_1^2 \\ &\quad + \lambda \sum_{j=2}^p (1-\alpha)\beta_j^2 + \alpha|\beta_j| \\ &= A\beta_1^2 + B\beta_1 + \lambda\alpha|\beta_1| + C\end{aligned}$$

$$\Rightarrow \hat{\beta}_1 = \frac{Q(\frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^* \beta^*), 2n\lambda\alpha)}{1 + 2n\lambda(1-\alpha)}, \quad \text{where}$$

$$Q(z, w) = \text{sgn}(z)(|z| - w)_+ = \begin{cases} z - w & z > 0, w < |z| \\ z + w & z < 0, w < |z| \\ 0 & w > z \end{cases}$$

Elastic Net Regression

- ▶ **Elastic Net:** $\tilde{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 + \lambda \cdot \sum_{j=1}^p (1 - \alpha) \beta_j^2 + \alpha |\beta_j|$.
- ▶ **Coordinate Descent:** Calculate $\tilde{\beta}$ for given λ .
- ▶ **Solution Path:** Repeat for $K \approx 100$ values of λ between λ_{\min} and λ_{\max} .



Picking the Value of λ

Objective 1: Minimize Prediction Error

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$

► **Parameter Estimation:** Elastic net for fixed α as a function of λ :

$$\tilde{\boldsymbol{\beta}}_{\lambda} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^p (1 - \alpha) \beta_j^2 + \alpha |\beta_j|.$$

► **Objective:** Suppose that $(y, \mathbf{x}) \sim f(y, \mathbf{x})$ have a joint distribution. Let $\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{obs}}$ denote the elastic net estimator for the given dataset $(\mathbf{y}_{\text{obs}}, \mathbf{X}_{\text{obs}})$. Then we wish to minimize the **mean square prediction error** (MSPE)

$$\hat{\lambda} = \arg \min_{\lambda} \text{MSPE}(\lambda), \quad \text{MSPE}(\lambda) = E[\{y - \mathbf{x}' \tilde{\boldsymbol{\beta}}_{\lambda}^{\text{obs}}\}^2].$$

Picking the Value of λ

Objective 1: Minimize Prediction Error

► **Model:** $y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \beta, \sigma^2)$

► **Parameter Estimation:** $\tilde{\beta}_\lambda = \arg \min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 + \lambda \cdot \sum_{j=1}^p (1 - \alpha) \beta_j^2 + \alpha |\beta_j|$.

► **Prediction Error:** $\hat{\lambda} = \arg \min_{\lambda} \text{MSPE}(\lambda)$, $\text{MSPE}(\lambda) = E[\{y - \mathbf{x}' \tilde{\beta}_\lambda^{\text{obs}}\}^2]$.

► **MSPE Estimation:** Use [cross-validation](#):

► Separate data into training and test sets: $(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ and $(\mathbf{y}_{\text{test}}, \mathbf{X}_{\text{test}})$.

► *MSPE estimate:* $\widehat{\text{MSPE}}(\lambda) = \sum_{i=1}^{n_{\text{test}}} \{y_i^{\text{test}} - \mathbf{x}_i^{\text{test}'} \tilde{\beta}_\lambda^{\text{train}}\}^2$,

where $\tilde{\beta}_\lambda^{\text{train}}$ is calculated from $(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}})$.

► *K-Fold CV:* (i) Randomly separate separate data into K sets (ii) MSPE estimate is

$$\widehat{\text{MSPE}}(\lambda) = \sum_{k=1}^K \widehat{\text{MSPE}}_k(\lambda),$$

where $\widehat{\text{MSPE}}_k(\lambda)$ has subset k as test set, and remaining data as training set.

Picking the value of λ

Objective 2: Don't miss any non-zero β_j 's

- ▶ **Model:** $y_i | x_i \stackrel{\text{ind}}{\sim} \mathcal{N}(x_i' \beta, \sigma^2)$
- ▶ **Covariance Test:** For each step k of LARS, can construct a test statistic T_{k+1} such that under

H_0 : All k non-zero β_j 's have been identified,

as $n, p \rightarrow \infty$ (but $p < n$) we have

$$T_{k+1} | H_0 \rightarrow \mathcal{F}(2, n - p).$$

- ▶ **In Practice:** Stop LARS at first value of $k - 1$ such that

```
pval <- pf(q = Tk, df1 = 2, df2 = n-p, lower.tail = FALSE)
```

is greater than 5%.

Resources

- ▶ **lars**: Package for LARS and Lasso.
- ▶ **covTest**: Implementation of covariance test. Paper by Lockhart et al (2014) can be found [here](#).
- ▶ **glmnet**: Very efficient LARS-type elastic net calculation for many GLMs.