# Statistics 101C - Week 5 - Model Selection in Regression

Shirong Xu

University of California, Los Angeles

shirong@stat.ucla.edu

October 28, 2024

# Regression

Given a data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is $p$-dimensional feature.

- Linear Regression framework:

$$\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)^T (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

- What if we include unrelated features?

# Regression

Given a data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is $p$-dimensional feature.

- Linear Regression framework:

$$\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)^T (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

- What if we include unrelated features?

# Regression

Given a data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is $p$-dimensional feature.

- Linear Regression framework:

$$\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)^T (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

- What if we include unrelated features?

# Example

```
C<-c(1,1,1,1,1,1,1,1)
x_1<-c(1,2,3,4,5,6,7,8)
x_2<-runif(8, -1, 1)
x_3<-runif(8, -1, 1)
x_4<-c(1,2,3,4,5,6,7,8.1)
y<-C + x_1 + runif(8, -1, 1)
Data <- data.frame(C,x_1,x_2,x_3,x_4,y)
model_1 <- lm(y ~ x_1, data = Data)
```

# Example: Model 1 $Y \sim X_1$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9586     0.5438   7.280 0.000342 ***
x_1           1.0149     0.1077   9.425 8.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.6979 on 6 degrees of freedom
Multiple R-squared:  0.9367,	Adjusted R-squared:  0.9262
F-statistic: 88.83 on 1 and 6 DF,  p-value: 8.111e-05
```

- $R^2$: 0.9367
- All features are significantly not equal to 0

# Example: Model 2 $Y \sim X_1 + X_2$

```
model_2 <- lm(y ~ x_1 + x_2, data = Data)
summary(model_2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.0059     0.5323   7.525 0.000656 ***
x_1           1.0067     0.1053   9.557 0.000212 ***
x_2          -0.5902     0.5178  -1.140 0.305998    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.6811 on 5 degrees of freedom
Multiple R-squared:  0.9498,	Adjusted R-squared:  0.9297 
F-statistic: 47.28 on 2 and 5 DF,  p-value: 0.0005652
```

- $R^2$: 0.9498 (increase by including an unrelated feature)

# Example: Model 2 $Y \sim X_1 + X_4$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.2160     0.5737   7.349 0.000732 ***
x_1          -9.3684     8.9618  -1.045 0.343731    
x_4          10.2975     8.8871   1.159 0.298906    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.6788 on 5 degrees of freedom
Multiple R-squared:  0.9501,    Adjusted R-squared:  0.9302 
F-statistic: 47.62 on 2 and 5 DF,  p-value: 0.0005556
```

- Both features $X_1$ and $X_4$ are significantly equal to 0.

# Sparse regression

Given training set $(\mathbf{x}_i, y_i)_{i=1}^n$ with $y_i \in \mathcal{R}$ and $\mathbf{x}_i \in \mathcal{R}^p$, it is assumed that

$$y_i = \beta_0 + \sum_{j=1}^{p_0} \beta_j x_{ij} + \epsilon_i,$$

where $p_0 \ll p$, and thus the sparsity.

# Sparse regression

Given training set $(\mathbf{x}_i, y_i)_{i=1}^n$ with $y_i \in \mathcal{R}$ and $\mathbf{x}_i \in \mathcal{R}^p$, it is assumed that

$$y_i = \beta_0 + \sum_{j=1}^{p_0} \beta_j x_{ij} + \epsilon_i,$$

where $p_0 \ll p$, and thus the sparsity.

- $\mathcal{A}^* = \{1, \ldots, p_0\}$ indexes the informative predictors, and $\{p_0 + 1, \ldots, p\}$ indexes the redundant predictors
- The goal of variable selection is to correctly detect $\mathcal{A}^*$ from $\{1, \ldots, p\}$
- We focus on linear regression models, while detecting nonlinear relationship is possible and largely open

# Why do we care?

- Multicollinearity: masked significance, inflated variance, ...

# Why do we care?

- Multicollinearity: masked significance, inflated variance, ...
- Prediction accuracy can be deteriorated due to overfitting when $p$ is large
- Interpretability can be unnecessarily complicated when irrelevant variables are included

# Popular techniques

- Best subset selection
  - Various information criteria, cross validation, ...
- Sequential variable selection
  - Forward/backward selection
- Shrinkage method
  - Lasso and its variants
- Dimension reduction
  - Principal component analysis, sufficient dimension reduction, ...

# Best subset selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors
2. For $k = 1, \ldots, p$
    a. Fit all $C_p^k$ models that contain exactly $k$ predictors
    b. Pick the best among these models and call it $\mathcal{M}_k$
3. Select a single best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$

# Best subset selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors
2. For $k = 1, \ldots, p$
   a. Fit all $C_p^k$ models that contain exactly $k$ predictors
   b. Pick the best among these models and call it $\mathcal{M}_k$
3. Select a single best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$

Popular selection criteria:

- Validation set
- Cross validation (CV) error
- "Estimate" test error by making an adjustment to the training error to account for overfitting

# Model selection criteria

For a linear model with $d$ predictors, denote its SSE as $SSE_d$,

- Mallow's $C_p$:

$$C_p = \frac{1}{n}(SSE_d + 2d\hat{\sigma}^2)$$

- Akaike information criterion (AIC):

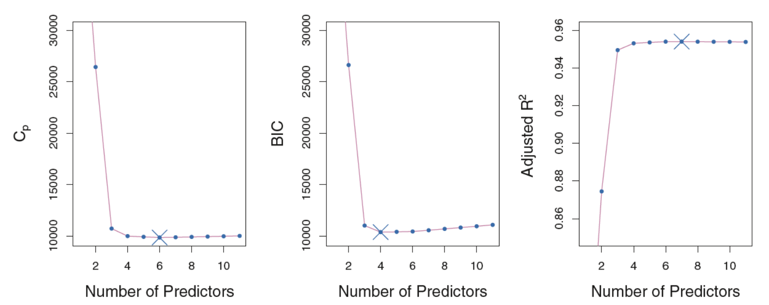
$$AIC = \frac{1}{n\hat{\sigma}^2}(SSE_d + 2d\hat{\sigma}^2)$$

- Bayesian information criterion (BIC):

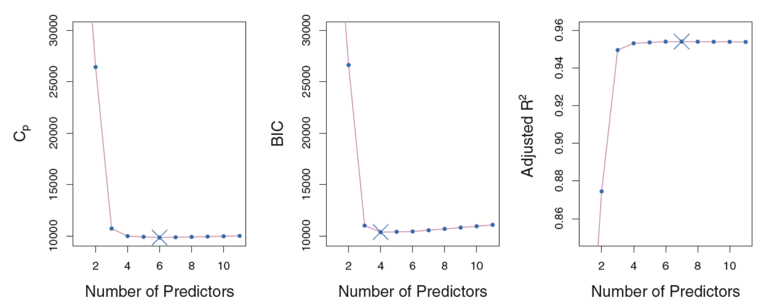
$$BIC = \frac{1}{n\hat{\sigma}^2}(SSE_d + \log(n)d\hat{\sigma}^2)$$

- Other criteria: other IC's, adjusted $R^2$

# An illustrative example



# An illustrative example




Question: Any drawbacks?

# Forward/backward selection

- Forward selection
  1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors
  2. For $k = 1, \ldots$
     a. Fit all models that contain $\mathcal{M}_{k-1}$ plus one additional predictor not in $\mathcal{M}_{k-1}$
     b. Pick the best among these models and call it $\mathcal{M}_k$
     c. Terminate if $\mathcal{M}_k$ is worse than $\mathcal{M}_{k-1}$ under certain model selection criterion

# Forward/backward selection

- Forward selection
  1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors
  2. For $k = 1, \ldots$
     a. Fit all models that contain $\mathcal{M}_{k-1}$ plus one additional predictor not in $\mathcal{M}_{k-1}$
     b. Pick the best among these models and call it $\mathcal{M}_k$
     c. Terminate if $\mathcal{M}_k$ is worse than $\mathcal{M}_{k-1}$ under certain model selection criterion
- Backward selection starts with $\mathcal{M}_p$ and iteratively delete predictors until the best model is found
- Stagewise selection mixes forward addition and backward deletion in each iteration

# Some remarks

- Forward/backward selection is computationally more efficient than subset selection
- It has no guarantee of the best possible model
- It usually performs well in practice
- Forward versus backward selection

# Shrinkage methods

- Shrinkage methods are formulated as

$$(\hat{\beta}_0, \hat{\beta}) = \underset{\beta}{\mathrm{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda J(\beta)$$

- Various choices of $J(\beta)$ lead to different shrinkage methods and possess different properties

# Shrinkage methods

- Shrinkage methods are formulated as

$$(\hat{\beta}_0, \hat{\beta}) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda J(\beta)$$

- Various choices of $J(\beta)$ lead to different shrinkage methods and possess different properties

- After centralization, it becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda J(\beta)$$

# Ridge regression

- Ridge regression uses an $L_2$-norm penalty, $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 = \beta^T \beta$,

$$\hat{\beta}_{\lambda}^{ridge} = \underset{\beta}{\operatorname{argmin}} \, (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \|\beta\|^2$$

# Ridge regression

- Ridge regression uses an $L_2$-norm penalty, $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 = \beta^T \beta$,

$$\hat{\beta}_\lambda^{ridge} = \underset{\beta}{\operatorname{argmin}} \, (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \|\beta\|^2$$

- The second term, $\lambda\|\beta\|^2$, is a shrinkage penalty, which shrinks the estimates of $\beta$ towards zero
- The tuning parameter $\lambda > 0$ controls the trade-off between regression fitting and coefficient shrinkage
- If $\lambda = 0$, ridge regression produces LSE; if $\lambda \to \infty$, estimates of $\beta$ will approach zero

# Ridge regression

- Solution of the ridge regression is

$$\hat{\beta}_{\lambda}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{y}$$

# Ridge regression

- Solution of the ridge regression is

$$\hat{\beta}_{\lambda}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^T \mathbf{y}$$

- An equivalent formulation,

$$\begin{aligned} \hat{\beta}_{\lambda}^{ridge} &= \operatorname*{argmin}_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &\quad \text{subject to } \|\beta\|^2 \leq s \end{aligned}$$

# Effective degree of freedom

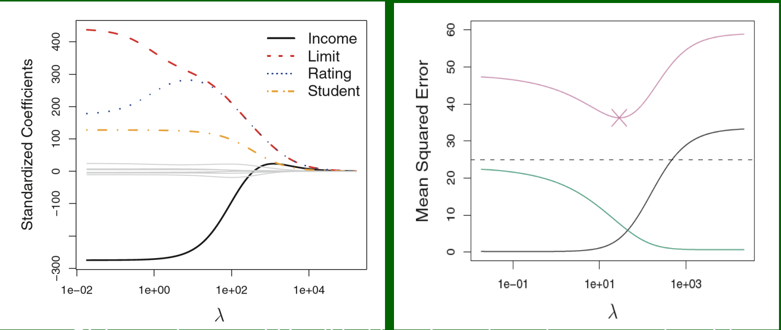
- The effective degree of freedom (df) of the ridge regression is

$$df(\hat{\mathbf{f}}_\lambda) = \mathrm{tr}\big(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\big) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

where $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$ is the SVD decomposition of $\mathbf{X}$, and $d_j$'s are the diagonal entries of $\mathbf{D}$

# An example

- In general, $\hat{\beta}_\lambda$ is a biased estimator that may have smaller MSE than the LSE estimator




Right panel: squared bias (black), variance (green), test error (purple)

# Lasso

- The lasso uses an $L_1$-norm penalty, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$,

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \; (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\|\beta\|_1$$

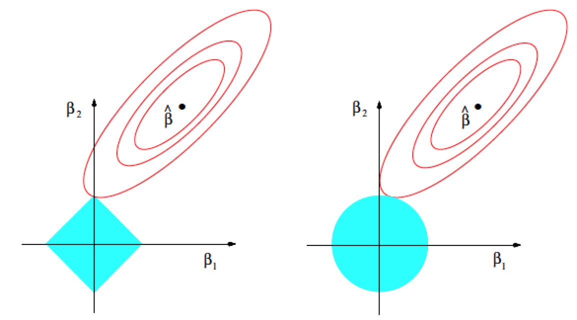
- Or equivalently,

$$\begin{aligned} \hat{\beta}^{lasso} &= \underset{\beta}{\operatorname{argmin}} \; (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \\ &\quad \text{subject to } \|\beta\|_1 \leq s \end{aligned}$$

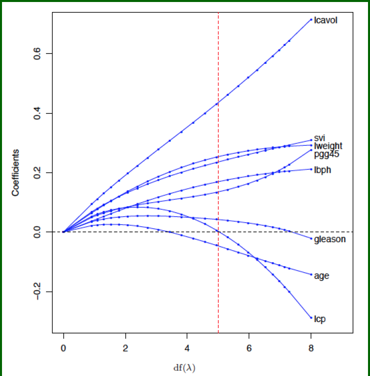
- No explicit solution in general, and a quadratic programming (QP) algorithm can be used to solve the optimization problem

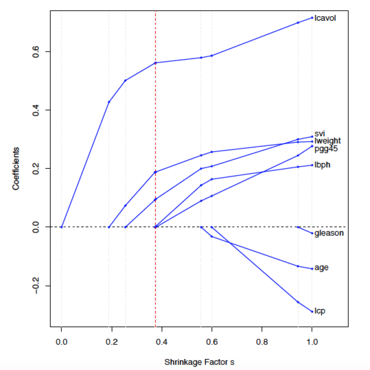
# Sparse solution

- Some coefficients of the lasso solution will become exactly zero, and thus it does some kind of continuous variable selection



# Example: Prostate cancer

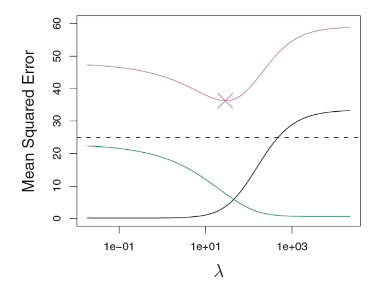


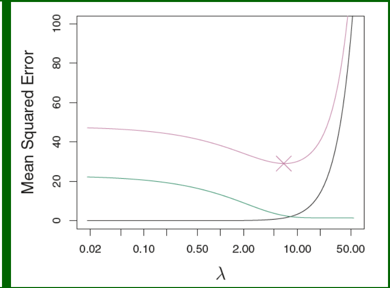




Left: ridge regression; Right: lasso regression

# Ridge vs Lasso

- Both lasso and ridge regression will shrink estimated coefficients while introducing some bias
- The lasso produces simpler and more interpretable models that involve only a subset of predictors
- It is unclear which one leads to better prediction accuracy in general though



# An orthogonal case

Consider a simple case with $n = p$ and $\mathbf{X} = \mathbf{I}_p$, then $\hat{\beta}_j^{ols} = y_j$,

- Ridge regression multiplies $\hat{\beta}_j^{ridge}$ by a constant, $\hat{\beta}_j^{ridge} = y_j/(1+\lambda)$

## An orthogonal case

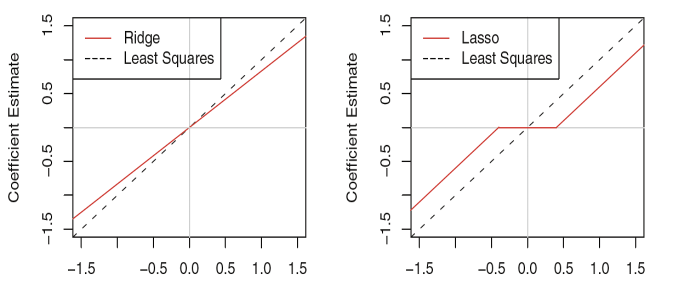
Consider a simple case with $n = p$ and $\mathbf{X} = \mathbf{I}_p$, then $\hat{\beta}_j^{ols} = y_j$,

- Ridge regression multiplies $\hat{\beta}_j^{ridge}$ by a constant, $\hat{\beta}_j^{ridge} = y_j/(1+\lambda)$
- Lasso truncates $\hat{\beta}_j^{ridge}$ towards zero by a constant, $\hat{\beta}_j^{lasso} = \mathrm{sign}(y_j)(|y_j| - \lambda/2)_+$

# An orthogonal case

Consider a simple case with $n = p$ and $\mathbf{X} = \mathbf{I}_p$, then $\hat{\beta}_j^{ols} = y_j$,

- Ridge regression multiplies $\hat{\beta}_j^{ridge}$ by a constant, $\hat{\beta}_j^{ridge} = y_j/(1+\lambda)$
- Lasso truncates $\hat{\beta}_j^{ridge}$ towards zero by a constant, $\hat{\beta}_j^{lasso} = \mathrm{sign}(y_j)(|y_j| - \lambda/2)_+$



# Bridge estimators

With $L_r(\beta) = \sum_{j=1}^{p} |\beta_j|^r$,

$$\hat{\beta}^{bridge} = \underset{\beta}{\operatorname{argmin}} \| \mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda L_r(\beta)$$

- $L_0(\beta) = \sum_{j=1}^{p} I(\beta_j \neq 0)$; (Hard thresholding)
- $L_1(\beta) = \sum_{j=1}^{p} |\beta_j|$; (Lasso)
- $L_2(\beta) = \sum_{j=1}^{p} \beta_j^2$; (Ridge regression)
- $L_\infty(\beta) = \max_j |\beta_j|$.

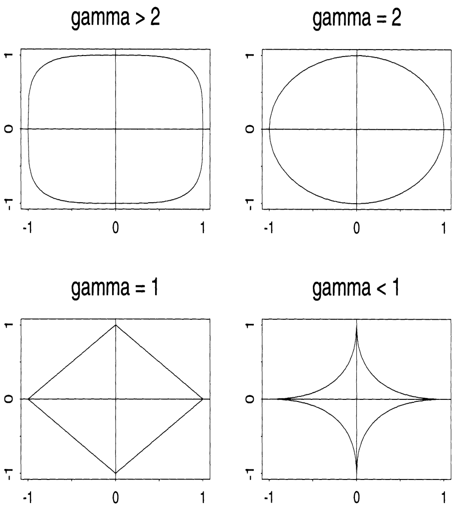



*Figure 1. Constrained Areas of Bridge Regressions with t = 1.*

# Nonnegative garrote

$$\min_{c} \; \frac{1}{2}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} c_j\hat{\beta}_j x_{ij}\Big)^2 + \lambda\sum_{j=1}^{p} c_j$$

subject to $c_j \geq 0$, and then $\hat{\beta}_j^{ng} = \hat{c}_j\hat{\beta}_j$.

# Nonnegative garrote

$$\min_{c} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} c_j\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} c_j$$

subject to $c_j \geq 0$, and then $\hat{\beta}_j^{ng} = \hat{c}_j\hat{\beta}_j$.

- The resulting estimator is

$$\hat{\beta}_j^{ng} = \left(1 - \frac{\lambda}{2\hat{\beta}_j^2}\right)_+ \hat{\beta}_j$$

- It is almost unbiased for large $|\hat{\beta}_j|$
- It shrinks small $|\hat{\beta}_j|$ to zero

# Other extensions

- Group lasso: if the $p$ variables are partitioned into $J$ groups, and then it is desirable to include or exclude the whole group

$$\min_{\beta} \; \frac{1}{2}\| \mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \sum_{j=1}^{J} \|\vec{\beta}_j\|_2,$$

where $\vec{\beta}_j$ is a coefficient vector for the $j$-th group

## Other extensions

- Group lasso: if the $p$ variables are partitioned into $J$ groups, and then it is desirable to include or exclude the whole group

$$\min_{\beta} \ \frac{1}{2}\| \mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \sum_{j=1}^{J} \|\vec{\beta}_j\|_2,$$

where $\vec{\beta}_j$ is a coefficient vector for the $j$-th group

- Elastic net:

$$\min_{\beta} \ \frac{1}{2}\| \mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$$

# Other extensions

- Group lasso: if the $p$ variables are partitioned into $J$ groups, and then it is desirable to include or exclude the whole group

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{J} \|\vec{\beta}_j\|_2,$$

where $\vec{\beta}_j$ is a coefficient vector for the $j$-th group
- Elastic net:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$$

- Fused lasso: penalize the difference between adjacent coef's

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=2}^{p} \|\beta_j - \beta_{j-1}\|_1,$$