Regularized regression I

Lasso and Ridge Regression

Regularization

Penalized regression models

- (Even) regression models can be over-parameterized (large p, small n)
- Shrinkage / Regularization methods
  - Consider model complexity in the estimation process by...
  - ...shrinking regression coefficients towards zero

$\rightarrow$ Ridge regression & Lasso

Ridge regression

OLS regression

$$\hat{\beta}_{OLS} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\}$$

$\rightarrow$ Minimizes ("only") RSS

Ridge regression

$$\begin{aligned}\hat{\beta}_{ridge} &= \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= RSS + \lambda \sum_{j=1}^{p} \beta_j^2\end{aligned}$$

$\rightarrow$ Introduces a **shrinkage penalty**: Fit - complexity trade-off

Ridge regression

OLS regression

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\}$$

→ Minimizes ("only") RSS

Ridge regression

$$\begin{aligned} \hat{\beta}_{ridge} &= \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \end{aligned}$$

→ Introduces a **shrinkage penalty**: Fit - complexity trade-off

Ridge regression

Comparing OLS and Ridge regression

$RSS = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$

$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

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

As a result...

- OLS regression needs $\mathbf{X}$ to be of full column rank
- Ridge regression (still) allows matrix inversion due to $\lambda\mathbf{I}$

Lasso

Other penalties are possible

Ridge regression

- Penalty on ℓ_2 norm of β
- $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Penalty on ℓ_1 norm of β
- $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$

Lasso

Other penalties are possible

Ridge regression

- Penalty on ℓ_2 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Penalty on ℓ_1 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

Penalty term

Increasing the penalty on model complexity

- $\lambda = 0$
  - Models are equivalent to OLS
- $\lambda \to \infty$
  - Ridge regression ($RSS + \lambda\|\beta\|_2^2$)
    - Coefficients are shrunken towards zero
    - Shrinks coefficients of correlated predictors towards each other
  - Lasso ($RSS + \lambda\|\beta\|_1$)
    - Coefficients are eventually shrunken exactly to zero (i.e. performs **variable selection**)
    - Erratic paths for correlated predictors

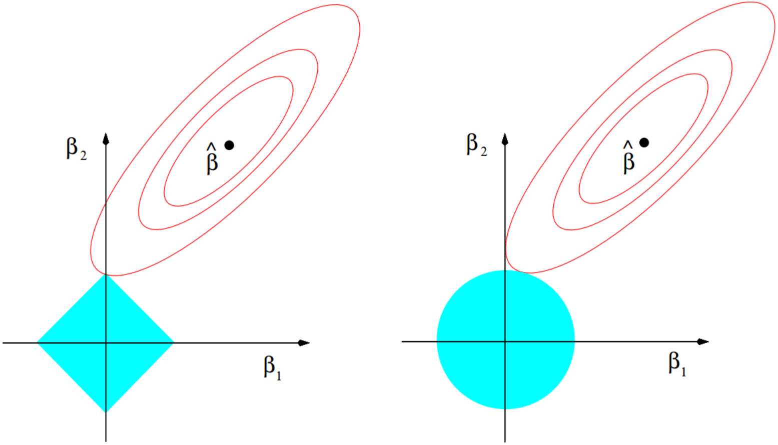
$\to$ The penalty λ is a tuning parameter

Penalty term

Figure: Constraint regions and RSS contours

(a) $|\beta_1| + |\beta_2| \leq t$

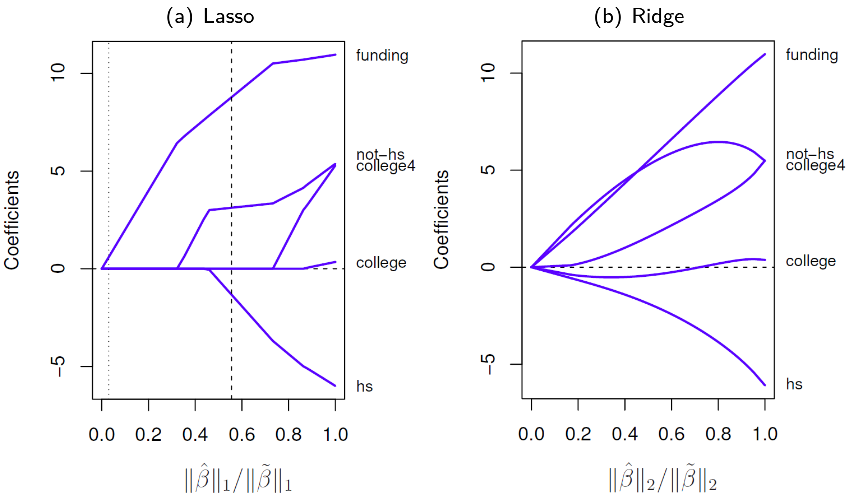
(b) $\beta_1^2 + \beta_2^2 \leq t^2$



Efron & Hastie (2016)

Penalty term

Figure: Coefficient paths



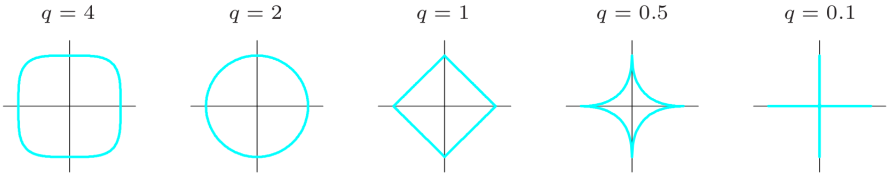
Hastie et al. (2015)

Penalty term

Other types of penalties?

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

Figure: Constraint regions for different values of q



Hastie et al. (2009)

Software Resources

Resources for R

- Standard package for ridge regression, lasso: `glmnet`
- Feature selection by filter: e.g. via `sbf` in `caret`
- Recursive feature elimination: e.g. `rfe` via in `caret`

References

Efron, B. and Hastie, T. (2016). Sparse Modeling and the Lasso. In Efron, B. and Hastie, T. (Eds.), *Computer Age Statistical Inference: Algorithms, Evidence and Data Science* (pp. 298–324). New York, NY: Cambridge University Press

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: Chapman & Hall/CRC.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.