Machine Learning Crash Course

Regularized Regression

Anthony Strittmatter

Literature

- James, Witten, Hastie, and Tibshirani (2013): "An Introduction to Statistical Learning", Springer, Chapter 6.2, download.
- Hastie, Tibshirani, and Friedman (2009): "Elements of Statistical Learning", 2nd ed., Springer, Chapter 3.4, download.

Best Subset Selection

- Consider we want to predict Y with a linear model including a constant and k predictors. Overall the data includes p covariates (excluding the constant). For the shake of illustration, assume $p = 100$.
- The number of possible models depends on k:
  - If $k = 0$, there is only one possible model.
  - If $k = 1$, there are 100 possible models.
  - If $k = 2$, there are 4,950 possible models.
  - If $k = 3$, there are 161,700 possible models.
  - If $k = 4$, there are 3,921,225 possible models.
- In general, the number of possible models for any k is (binomial expansion)

$$\binom{p}{k} = \frac{p!}{k!(p-k)!},$$

  or 2^p models across all possibel k's.
- Select optimal k using cross-validation.

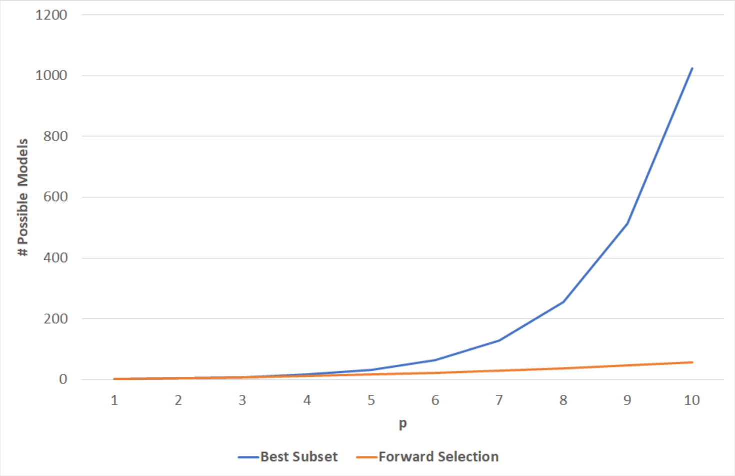
Forward Stepwise Selection

- Impose a bottom-up hierarchical structure on the covariates:
  - The first model ($k = 0$) contains only a constant.
  - The second model ($k = 1$) adds to the constant one out of p possible covariates.
  - The third model ($k = 2$) equals the second model, but adds one out of $p - 1$ possible covariates.
  - The fourth model ($k = 3$) equals the third model, but adds one out of $p - 2$ possible covariates.
- In general, the number of possible models is

$$1 + \frac{p(p+1)}{2}.$$

- Select optimal k using cross-validation.

Number of Possible Models



Ridge

Summation notation:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

where $\lambda \geq 0$ is the penalty parameter and the number of covariates p can be high-dimensional ($p \gg N$).

→ Note that coefficient size depends on the scaling of x_j. It is best practice to standardise non-binary x_j. In the following, we assume that all covariates are standardized.

Matrix notation:

$$\min_{\beta}\left\{(Y-X\beta)'(Y-X\beta)+\lambda\|\beta\|_2^2\right\}$$

where $\beta=(\beta_1,\beta_2,\cdots\beta_p)'$ does not include the constant term $\beta_0=\frac{1}{N}\sum_{i=1}^{N}y_i$. The squared l_2-norm is $\|\beta\|_2^2=\beta'\beta=\sum_{j=1}^{p}\beta_j^2$.

First Order Condition

Partial derivative w.r.t. β:

$$-2X'(Y - X\beta) + 2\lambda I\beta = 0$$

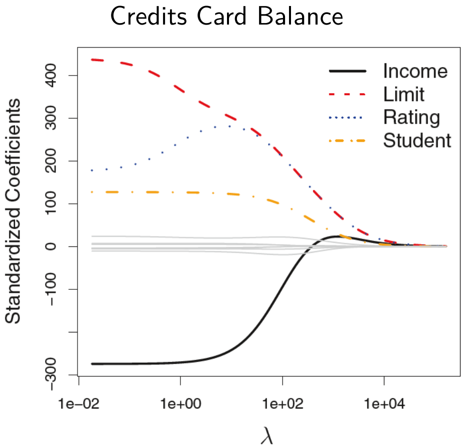
where I is a $p \times p$ identity matrix.

Closed-form solution:

$$\widehat{\beta} = (X'X + \lambda I)^{-1}X'Y$$

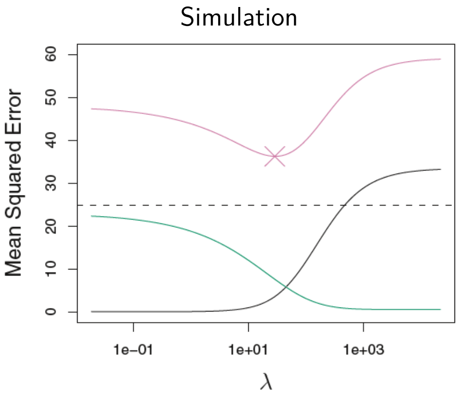
with $(X'X + \lambda I)$ being positive definite.

Ridge Coefficients



Source: James et al. (2013), p. 216

Ridge: Variance-Bias Trade-Off



Note: squared bias (black), variance (green), MSE (red)

Source: James et al. (2013), p. 218

Selection of Optimal Penalty Parameter

Cross-Validation (CV) Algorithm

Firewall Principle

Why do we use the hold-out-sample to evaluate the prediction power?

- If we try many tuning parameter values, we may end up overfitting even in cross-validation samples.
- The cross-validation performance is an aggregation over multiple different prediction functions, which differs from the single prediction function we finally estimate.

Lasso

Summation notation:

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

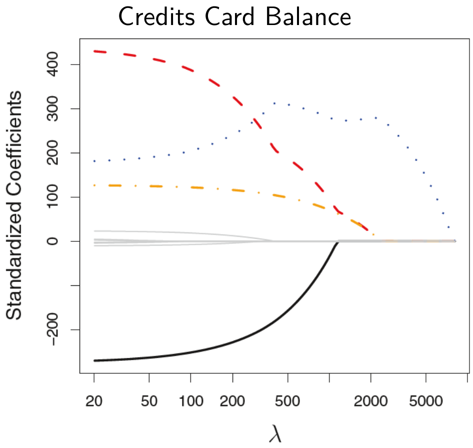
where $\lambda \geq 0$ is the penalty parameter.

Matrix notation:

$$\min_{\beta} \{ (Y - X\beta)'(Y - X\beta) + \lambda \|\beta\|_1 \}$$

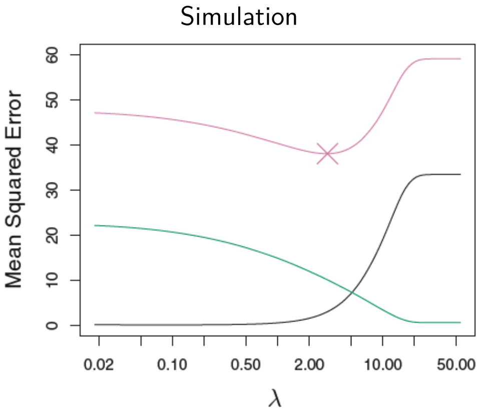
with $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (l_1-norm).

Lasso Coefficients



Source: James et al. (2013), p. 220

Lasso: Variance-Bias Trade-Off



Note: squared bias (black), variance (green), MSE (red)

Source: James et al. (2013), p. 223

Constrained Regression

- OLS residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

- Penalized regression:
  - Lagrangian operator

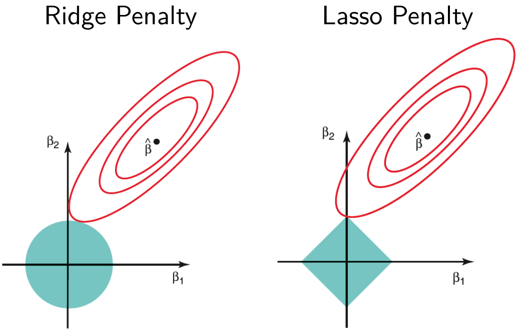
$$\min_{\beta}\{RSS + \lambda \sum_{j=1}^{p} p(\beta_j)\}$$

  - Constrained regression

$$\min_{\beta}\{RSS\} \text{ s.t. } \sum_{j=1}^{p} p(\beta_j) \leq c$$

where $p(\beta_j) = \beta_j^2$ for Ridge and $p(\beta_j) = |\beta_j|$ for Lasso.

Constraint Regions



Source: James et al. (2013), p. 222

Simple Example

- Consider $X = I$ with dimension $p = N$.
- OLS model
$$\sum_{j=1}^{p}(y_j - \beta_j)^2,$$
such that the estimated OLS coefficients are $\widehat{\beta}_j = y_j$.
- Ridge model
$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$
such that the estimated Ridge coefficients are $\widehat{\beta}_j^R = \widehat{\beta}_j/(1 + \lambda)$.

Simple Example (cont.)

- LASSO model

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

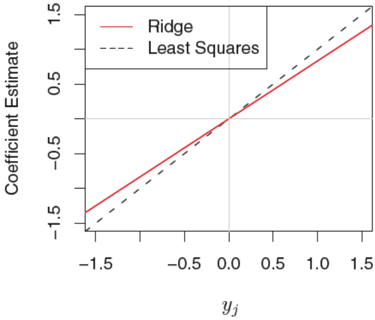
such that the estimated LASSO coefficients are

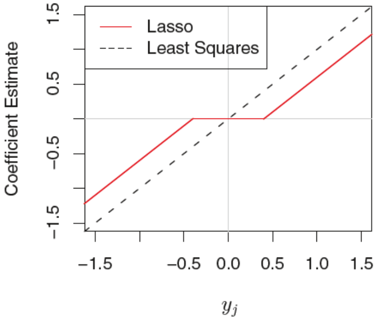
$$\widehat{\beta}_j^L = \begin{cases} \widehat{\beta}_j - \lambda/2 & \text{if } \widehat{\beta}_j > \lambda/2, \\ \widehat{\beta}_j + \lambda/2 & \text{if } \widehat{\beta}_j < -\lambda/2, \\ 0 & \text{if } |\widehat{\beta}_j| \leq \lambda/2, \end{cases}$$

which corresponds to the soft-thresholding operator

$$\widehat{\beta}_j^L = sign(\widehat{\beta}_j)(|\widehat{\beta}_j| - \lambda/2)_+$$

Simple Example (cont.)





Source: James et al. (2013), p. 226

Coordinate Descent Algorithm for Lasso

$$\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda_s\sum_{j=1}^{p}|\beta_j|\right\}$$

(1) Specify a grid of $s = 1, ...S$ tuning parameters $\lambda_s \in \{\lambda_1, \lambda_2, ..., \lambda_S\}$

(2) Take residuals $y_i^* = y_i - \frac{1}{N}\sum_{i=1}^{N} y_i$ and initialise $\beta_j = 0$

(3) Circulate repeatedly over all $j = 1, ..., p$ until convergence:

  (a) Compute the partial residuals by $r_{ij} = y_i^* - \sum_{k\neq j} x_{ik}\beta_k$
  (b) Calculate the simple univariate OLS coefficient $\tilde{\beta}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}r_{ij}$
  (c) Update β_j with the soft-thresholding operator:

$$\beta_j = sign(\tilde{\beta}_j)(|\tilde{\beta}_j| - \lambda_s)_+$$

(4) Repeat (3) for $s = 1, ..., S$

Note: Standardisation of x is required

Post-Lasso

- Coefficients of LASSO $\widehat{\beta}_j$ are biased when $\lambda > 0$, because the penalty terms shrinks the coefficients towards zero.
- Post-LASSO enables an easy interpretation.
- **Idea:**
  1. Estimate a Lasso model with the cross-validated optimal penalty.
  2. Estimate an OLS model (called Post-Lasso) which includes all variables with non-zero coefficients from the first-step Lasso model.
- **Problems:**
  - Post-Lasso coefficients are also biased in the presence of omitted variable bias.
  - The first-step model selection of the Lasso is often unstable.

Other Extensions

Elastic Net:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}(\alpha|\beta_j|+(1-\alpha)\beta_j^2)\right\}$$

Best Subset Selection:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}1\{\beta_j\neq 0\}\right\}$$