Shrinkage Methods for Linear Models

Bin Li

IIT Lecture Series

Outline

- Linear regression models and least squares
- Subset selection
- Shrinkage methods
    - Ridge regression
    - The lasso
    - Subset, ridge and the lasso
    - Elastic net
- Shrinkage method for classification problem.

Linear Regression & Least Squares

- Input vector: $X^T = (X_1, X_2, \ldots, X_p)$
- Real valued output vector: $Y = (y_1, y_2, \ldots, y_n)$
- Linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- *Least squares* estimate minimizes

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{p} X_j \beta_j))^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

Least Square Estimate

- Differentiate $RSS(\beta)$ w.r.t. β:

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \qquad (1)$$

$$\frac{\partial^2 RSS}{\partial \beta \beta^T} = +2\mathbf{X}^T\mathbf{X} \qquad (2)$$

- Assume **X** has full column rank, (2) is positive definite.
- Set first derivative to zero:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \Longrightarrow \hat{\beta}^{ols} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- Variance of $\hat{\beta}^{ols}$: $Var(\hat{\beta}^{ols}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

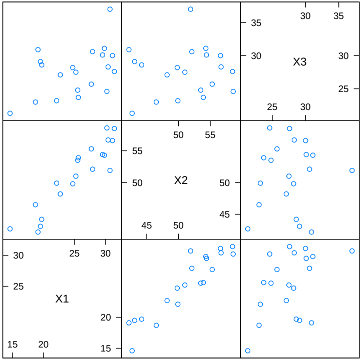
Gauss-Markov Theorem

- Suppose we have: $Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i$
- We need the following assumptions:
  - $E(\epsilon_i) = 0, \ \forall i$
  - $Var(\epsilon_i) = \sigma^2, \ \forall i$
  - $Cov(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j$
- Gauss-Markov Theorem says that:
  ordinary least squares estimator (OLS) is the best linear unbiased estimator (BLUE) of β.
  - $E(\hat{\beta}^{ols}) = \beta$
  - Among all the linear unbiased estimators $\hat{\beta}$, OLS estimator has the minimum

$$\sum_{j=1}^{p} E[(\hat{\beta}_j - \beta_j)^2] = \sum_{j=1}^{p} Var(\hat{\beta}_j)$$

Body Fat Example

This is a study of the relation of amount of body fat (Y) to several possible X variables, based on a sample of 20 healthy females 25-34 years old. There are three X variables: triceps skinfold thickness (X_1); thigh circumference (X_2); and midarm circumference (X_3).



Scatter Plot Matrix

Body Fat Example (cont)

```
fit<-lm(Y~.,data=bfat)
summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-Squared: 0.8014, Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Remarks

- Estimation:
  - OLS is unbiased but when some input variables are highly correlated, $\hat{\beta}$'s may be highly variable.
  - To see if a linear model $\hat{f}(\mathbf{x}) = \mathbf{x}^T\hat{\beta}$ is a good candidate, we can ask ourselves two questions:
    1. Is $\hat{\beta}$ close to the true β?
    2. Will $\hat{f}(\mathbf{x})$ fit future observations well?
  - Bias-variance trade-off.
- Interpretation:
  - Determine a smaller subset that exhibit the strongest effects
  - Obtain a "big picture" by sacrifice some small details

Is $\hat{\beta}$ close to the true β?

- To answer this question, we might consider the **mean squared error** of our estimate $\hat{\beta}$:
  - Squared distance of $\hat{\beta}$ to the true β:

$$\begin{aligned}
MSE(\hat{\beta}) &= \mathbb{E}[||\hat{\beta} - \beta||^2] = \mathbb{E}[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] \\
&= \mathbb{E}[(\hat{\beta} - \mathbb{E}\hat{\beta} + \mathbb{E}\hat{\beta} - \beta)^T(\hat{\beta} - \mathbb{E}\hat{\beta} + \mathbb{E}\hat{\beta} - \beta)] \\
&= \mathbb{E}[||\hat{\beta} - \mathbb{E}\hat{\beta}||^2] + 2\mathbb{E}[(\hat{\beta} - \mathbb{E}\hat{\beta})^T(\mathbb{E}\hat{\beta} - \beta)] + ||\mathbb{E}\hat{\beta} - \beta||^2 \\
&= \mathbb{E}[||\hat{\beta} - \mathbb{E}\hat{\beta}||^2] + ||\mathbb{E}\hat{\beta} - \beta||^2 \\
&= \mathrm{tr}(\mathrm{Variance}) + ||\mathrm{Bias}||^2
\end{aligned}$$

- For OLS estimate, we have

$$MSE(\hat{\beta}) = \sigma^2 \mathrm{tr}[(\mathbf{x}^T\mathbf{x})^{-1}]$$

Will $\hat{f}(\mathbf{x})$ fit future observation well?

- Just because $\hat{f}(\mathbf{x})$ fits our data well, this doesn't mean that it will be a good fit to a new dataset.
- Assume $y = f(\mathbf{x}) + \epsilon$ where $\mathbb{E}\epsilon = 0$ and $\text{Var}(\epsilon) = \sigma_\epsilon^2$
- Denote $\mathcal{D}$ as the training data to fit the model $\hat{f}(x, \mathcal{D})$
- **Expected prediction error** (squared loss) at an input point $\mathbf{x} = \mathbf{x}^{new}$ can be decomposed as

$$\begin{aligned}
\text{Err}(x^{new}) &= \mathbb{E}_{\mathcal{D}}\left[(Y - \hat{f}(X, \mathcal{D}))^2 | X = x^{new}\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[(Y - f(X))^2 | X = x^{new}\right] \\
&\quad + \mathbb{E}_{\mathcal{D}}\left[(f(X) - \mathbb{E}_{\mathcal{D}}\hat{f}(X, \mathcal{D}))^2 | X = x^{new}\right] \\
&\quad + \mathbb{E}_{\mathcal{D}}\left[(\mathbb{E}_{\mathcal{D}}\hat{f}(X, \mathcal{D}) - \hat{f}(X, \mathcal{D}))^2 | X = x^{new}\right] \\
&= \sigma_\epsilon^2 + \text{Bias}^2\{\hat{f}(x^{new})\} + \text{Var}\{\hat{f}(x^{new})\} \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

Subset Selection

- Best-subset selection
  - an efficient algorithm *leaps and bounds* procedure - Furnival and Wilson (1974)
  - feasible for p as large as 30 to 40.
- Forward- and backward-stepwise selection
  - FS selection is a *greedy* algorithm producing a nested sequence of model
  - BS selection sequentially deletes predictor with the least impact on the fit
- Two-way stepwise-selection strategies
  - consider both *forward* and *backward* moves at each step
  - **step** function in **R** package uses AIC criterion

Ridge Regression – A Shrinkage Approach

- Shrinks the coefficients by imposing a constraint.

$$\hat{\beta}^R = \arg\min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq C$$

- Equivalent *Lagrangian form*:

$$\hat{\beta}^R = \arg\min_{\beta} RSS + \lambda \sum_{j=1}^{p} \beta_j^2, \quad \lambda \geq 0.$$

- Solution of ridge regression:

$$\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Ridge Regression (cont'd)

- Apply correlation transformation on X and Y
- The estimated standardized regression coefficients in ridge regression is:
$$\hat{\beta}^R = (\mathbf{r}_{XX} + \lambda \mathbf{I})^{-1}\mathbf{r}_{YX}$$
    - $\mathbf{r}_{XX}$ is the correlation matrix of $\mathbf{X}$ (i.e. $\mathbf{X}^T\mathbf{X}$ on transformed $\mathbf{X}$)
    - $\mathbf{r}_{YX}$ is the correlation matrix between $\mathbf{X}$ and Y (i.e. $\mathbf{X}'\mathbf{y}$ on transformed $\mathbf{X}$ and $\mathbf{y}$)
- λ is a tuning parameter in ridge regression
    - λ reflects the amount of bias in the estimators. $\lambda \uparrow \Rightarrow$ bias $\uparrow$
    - λ stabilize the regression coefficient β. $\lambda \uparrow \Rightarrow$ *VIF* $\downarrow$
- Three ways to choose λ
    - Based on *ridge trace* of $\hat{\beta}^R$ and *VIF*
    - Based on *external validation* or *cross-validation*
    - Use generalize cross-validation.

Correlation Transformation and Variance Inflation Factor

- Correlation transformation is used to standardized variables:

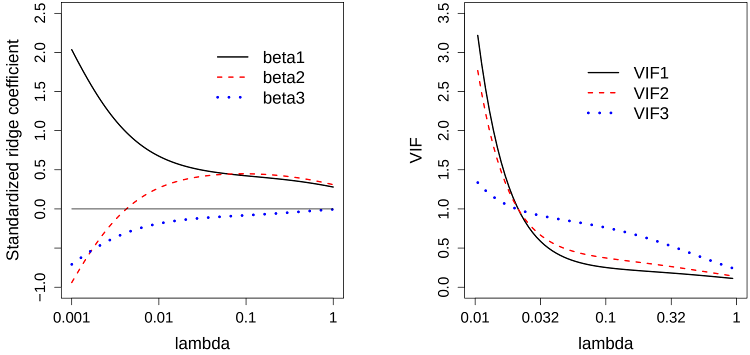
$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \quad \text{and} \quad X_{ij}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ij} - \bar{X}_j}{s_{X_j}}\right)$$

- For OLS, the VIF_j for the j^{th} variable is defined as the j^{th} diagonal element of the matrix $\mathbf{r}_{XX}^{-1}$.
- It is equal to $1/(1-R_j^2)$, where R_j^2 is the R-square when X_j is regressed on the rest $p-1$ variables in the model.
- Hence, when VIF_j is 1, it indicates X_j is not linearly related to the other X's. Otherwise VIF_j will be greater than 1.
- In ridge regression, the VIF values are the diagonal elements of the following matrix:

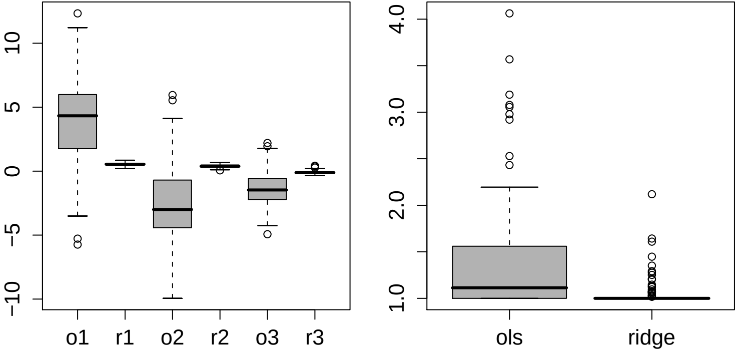
$$(\mathbf{r}_{XX} + \lambda \mathbf{I})^{-1}\, \mathbf{r}_{XX}\, (\mathbf{r}_{XX} + \lambda \mathbf{I})^{-1}$$

Body Fat Example (cont'd)

- When λ is around 0.02, VIF_j are all close to 1
- When λ reaches 0.02, $\hat{\beta}^R$ stabilize.
- Note: horizontal axis is not equally spaced (in log scale)



Body Fat Example (cont'd)



$Var\{\hat{\beta}\}$ for OLS (ridge): 13.9 (0.016), 11.1 (0.015), 2.11 (0.019)
Average of MSE in OLS and ridge: 0.0204 and 0.0155.

Bias and Variance of Ridge Estimate

$$\begin{aligned}
\mathbb{E}\hat{\beta}^R &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T(\mathbf{X}\beta+\epsilon)\right] \\
&= \left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta \\
&= \beta-\lambda\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\beta
\end{aligned}$$

Hence, the ridge estimate $\hat{\beta}^R$ is biased. The bias is:

$$\mathbb{E}\hat{\beta}^R-\beta=-\lambda\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\beta$$

$$\begin{aligned}
Var(\hat{\beta}^R) &= Var\left\{\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T(\mathbf{X}\beta+\epsilon)\right\} \\
&= Var\left\{\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\epsilon\right\} \\
&= \left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\sigma^2
\end{aligned}$$

Bias-variance tradeoff

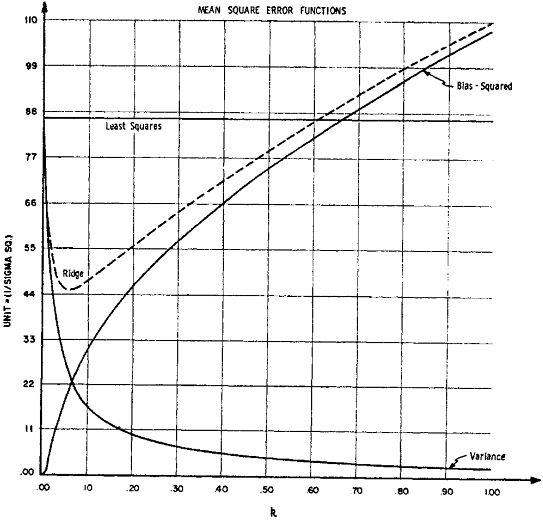


Figure from Hoerl and Kennard, 1970.

Data Augmentation to Solve Ridge

$$\begin{aligned}
\hat{\beta}^R &= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta)^2 + \sum_{j=1}^{p} (0 - \sqrt{\lambda}\beta_j)^2 \right\} \\
&= \arg\min_{\beta} \left\{ \sum_{i=1}^{n+p} (y_i^{\lambda} - \mathbf{z}_i^{\lambda}\beta)^2 \right\} = (\mathbf{y}_{\lambda} - \mathbf{z}_{\lambda}\beta)^T (\mathbf{y}_{\lambda} - \mathbf{z}_{\lambda}\beta),
\end{aligned}$$

where

$$\mathbf{z}_{\lambda} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \\ \sqrt{\lambda} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda} & \cdots & 0 \\ 0 & 0 & \cdots & \sqrt{\lambda} \end{bmatrix} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_p \end{pmatrix} \text{ and } \mathbf{y}_{\lambda} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Lasso

Lasso is a loop of rope that is designed to be thrown around a target and tighten when pulled. Figure is from Wikipedia.org.

LASSO: Least Absolute Shrinkage and Selection Operator

- Shrinks the coefficients by imposing the constraint:

$$\hat{\beta}^R = \arg\min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq C$$

- Equivalent *Lagrangian form*:

$$\hat{\beta}^R = \arg\min_{\beta} RSS + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda \geq 0.$$

- When C is sufficiently small (or λ is sufficiently large), some of the coefficients will be **exactly** zero. Lasso does subset selection in a *continuous* way.

Prostate Cancer Example

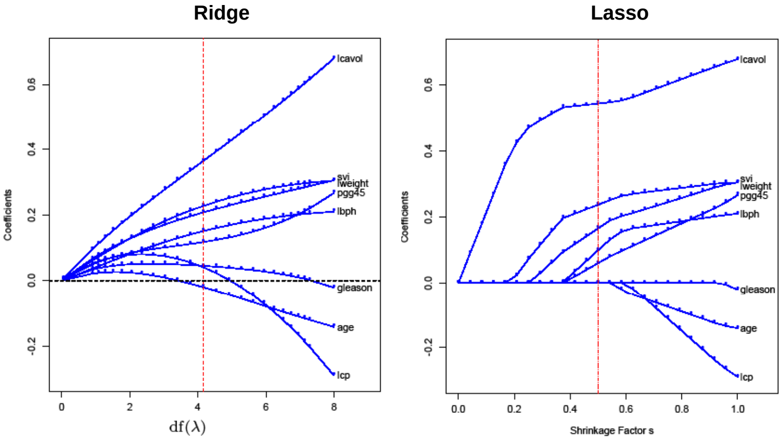
- Eight input variables:
  - Log caner volume (lcavol)
  - log prostate weight (lweight)
  - Age (age)
  - log of benigh prostatic hyperplasia (lbph)
  - Seminal vesicle invasion (svi)
  - Log capsular penetration (lcp)
  - Gleason score (gleason)
  - Percent of Gleason scores 4 or 5 (pgg45)
- Response variable: log of prostate-specific antigen (lpsa)
- Sample size: 97 subjects

Estimated Coefficients in Prostate Cancer Example

- Training set: 67 obs.; testing set: 30 obs.
- Ten-fold cross-validation to determine λ in ridge and lasso
- Following table is from HTF 2009

Term	OLS	Best Subset	Ridge	Lasso	PCR
Intercept	2.465	2.477	2.452	2.468	2.497
lcavol	0.680	0.740	0.420	0.533	0.543
lweight	0.263	0.316	0.238	0.169	0.289
age	-0.141	0	-0.046	0	-0.152
lbph	0.210	0	0.162	0.002	0.214
svi	0.305	0	0.227	0.094	0.315
lcp	-0.288	0	0.001	0	-0.051
gleason	-0.021	0	0.040	0	0.232
pgg45	0.267	0	0.133	0	-0.056
Test error	0.521	0.492	0.492	0.279	0.449

Solution Path for Ridge and Lasso



$$df(\lambda) = trace[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T] \quad s = \sum_{j=1}^{8} |\hat{\beta}^{lasso}| / \sum_{j=1}^{8} |\hat{\beta}^{ols}|$$

Figure from HTF 2009.

Best Subset, Ridge and Lasso

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \times I[rank(\lvert\hat{\beta}_j\rvert) \leq M]$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$sign(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

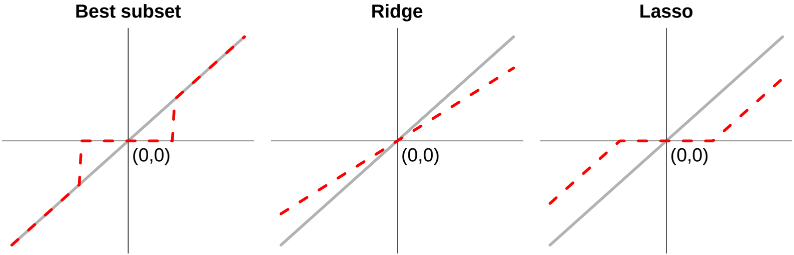


Figure from HTF 2009.

Geometric View of Ridge and Lasso

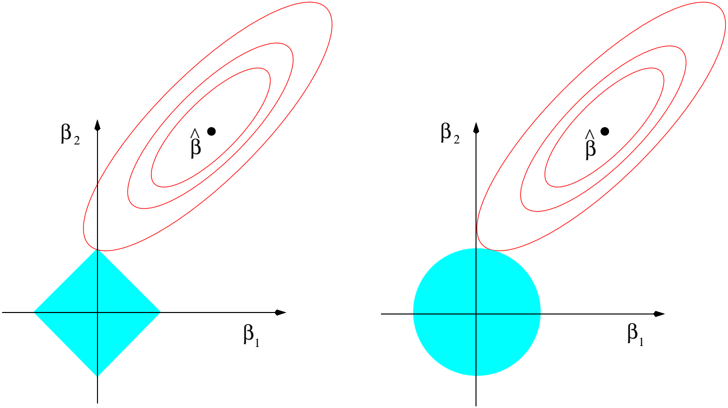


Figure from HTF 2009.

Comparison of Prediction Performance

- Model: $y = \beta^T x + \sigma\epsilon$, where $\epsilon \sim N(0,1)$.
- Training set: 20; test set: 200; 50 replications.
- Simulation 1: $\beta = 0.85 \times (1,1,1,1,1,1,1,1)^T$, $\sigma = 3$

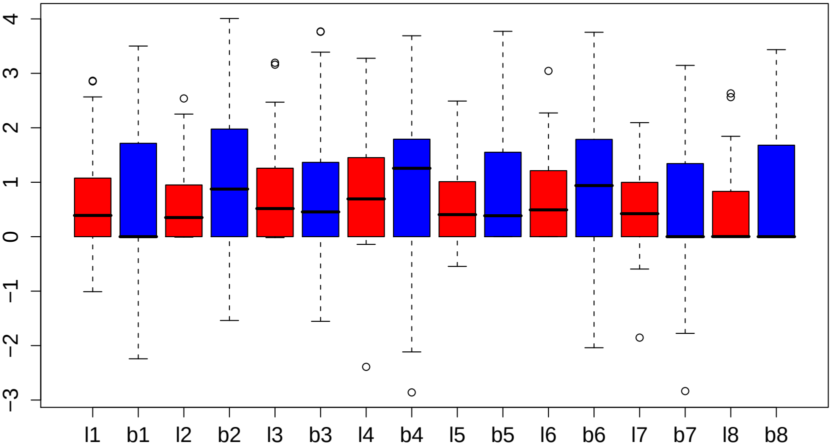
Method	Median MSE	Avg. # of 0's
OLS	6.50 (0.64)	0.0
Lasso	4.87 (0.35)	2.3
Ridge	2.30 (0.22)	0.0
Subset	9.05 (0.78)	5.2

- Simulation 2: $\beta = (5,0,0,0,0,0,0,0)^T$, $\sigma = 2$

Method	Median MSE	Avg. # of 0's
OLS	2.89 (0.04)	0.0
Lasso	0.89 (0.01)	3.0
Ridge	3.53 (0.05)	0.0
Subset	0.64 (0.02)	6.3

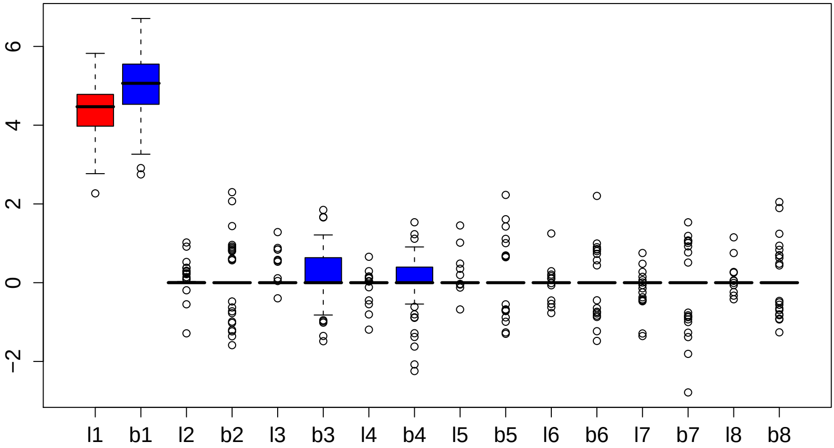
Both tables are from Tibshirani (1996).

Stability in $\hat{\beta}$ in Simulation 1



$$\beta = 0.85 \times (1,1,1,1,1,1,1,1)^T$$

Stability in $\hat{\beta}$ in Simulation 2

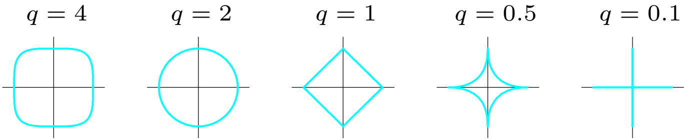


$$\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$$

Other Issues

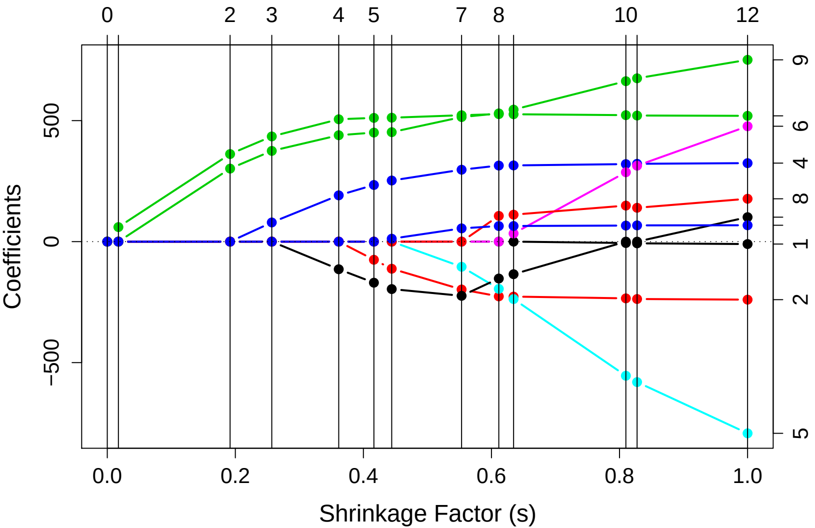
- Bridge regression family

$$\hat{\beta}^B = \arg\min_{\beta} RSS + \lambda \sum_{j=1}^{p} |\beta_j|^q, \quad q > 0.$$



- Best subset selection
- Bayesian interpretation
- Computation of Lasso & piecewise linear solution path

Piecewise Linear Solution Path in Lasso



Limitations of LASSO

- In the $p > n$ case, the lasso selects at most n variables before it saturates. The number of selected variables is bounded by the number of samples. This seems to be a limiting feature for a variable selection method.
- If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.
- For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).

Elastic Net Regularization

$$\hat{\beta}^{ENET} = \arg\min_{\beta} RSS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

- The L_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the L_1 regularization path.
- The elastic net objective function can be expressed as:

$$\hat{\beta}^{ENET} = \arg\min_{\beta} RSS + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

A Simple Illustration: ENET vs. LASSO

- Two independent "hidden" factors z_1 and z_2

$$z_1 \sim U(0, 20) \quad z_2 \sim U(0, 20)$$

- Response vector $y = z_1 + 0.1 z_2 + N(0, 1)$
- Suppose only observe predictors:

$$\begin{aligned} x_1 &= z_1 + \epsilon_1, \quad x_2 = -z_1 + \epsilon_2 \qquad x_3 = z_1 + \epsilon_3 \\ x_4 &= z_2 + \epsilon_4, \quad x_5 = -z_2 + \epsilon_5 \qquad x_6 = z_2 + \epsilon_6 \end{aligned} \tag{3}$$

- Fit the model on $(x_1, x_2, x_3, x_4, x_5, x_6, y)$.
- An "oracle" would identify x_1,x_2,x_3 (the z_1 group) as the most important variables.

Simulation Studies

Simulation example 1: 50 data sets consisting of 20/20/200 observations and 8 predictors. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. $\text{cor}(\mathbf{x}_i, \mathbf{x}_j) = (0.5)^{|i-j|}$.

Simulation example 2: Same as example 1, except $\beta_j = 0.85$ for all j.

Simulation example 3: 50 data sets consisting of 100/100/400 observations and 40 predictors. $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$ and $\sigma = 15$; $\text{cor}(x_i, x_j) = 0.5$ for all i, j.

Simulation example 4: 50 data sets consisting of 50/50/400 observations and 40 predictors. $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$ and $\sigma = 15$.

$$\begin{aligned}
\mathbf{x}_i &= Z_1 + \epsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1, \cdots, 5, \\
\mathbf{x}_i &= Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, \cdots, 10, \\
\mathbf{x}_i &= Z_3 + \epsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11, \cdots, 15, \\
\mathbf{x}_i &\sim N(0,1), \qquad \mathbf{x}_i \quad \text{i.i.d} \qquad i = 16, \ldots, 40.
\end{aligned}$$

Simulation Results

Median MSE for the simulated examples

Method	Ex.1	Ex.2	Ex.3	Ex.4
Ridge	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic Net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
No re-scaling	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

Variable selection results

Method	Ex.1	Ex.2	Ex.3	Ex.4
Lasso	5	6	24	11
Elastic Net	6	7	27	16

Other Issues

- Elastic net with scaling correction: $\hat{\beta}_{enet} = (1 + \lambda_2)\hat{\beta}^{ENET}$.
- Keep the grouping effect and overcome the double shrinkage by the quadratic penalty (too much shrinkage/bias towards zero).
- The elastic net solution path is also *piecewise linear*.
- Solviing elastic net is essentially solving a LASSO problem with augmented data.
- Coordinate descent algorithm efficiently solves the elastic net solution.
- Elastic net can also be used in classification and other problems like GLM.
- The *glmnet* package in *R* use coordinate descent algorithm solves elastic-net type problems.

South African heart disease

- A subset of Coronary Risk-Factor Study (CORIS) survey.
- 462 white males between 15 and 64 from Western Cape, South Africa.
- Response variable: presence (160) or absence (302) of myocardial infarction.
- Seven input variables: systolic blood pressure (sbp), obesity, tobacco, ldl, famhist, alcohol, age. Table below is from EOSL (2009).

Variable	Coef	SE	Z score
Intercept	-4.130	0.964	-4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	-1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

Solution path for L_1 regularized logistic regression

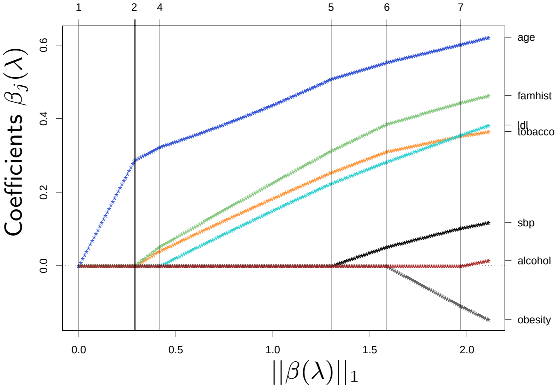


FIGURE 4.13. *L_1 regularized logistic regression coefficients for the South African heart disease data, plotted as a function of the L_1 norm. The variables were all standardized to have unit variance. The profiles are computed exactly at each of the plotted points.*

Figure from EOSL, 2009.

Reference

- "Ridge regression: biased estimation for nonorthogonal problems." Hoerl and Kennard (1970). *Technometrics*, **12**, 55-67.
- "A statistical view of some chemometrics regression tools." Frank and Friedman (1993). *Technometrics*, **35**, 109-135.
- "Regression shrinkage and selection via the lasso." Tibshirani (1996). *JRSSB*, **58**, 267-288.
- "Regularization and Variable Selection via the Elastic Net". Zou and Hastie (2005). *JRSSB*,**67**, 301320.
- "Elements of Statistical Learning." 2nd ed. Hastie, Tibshirani and Friedman (2009), Springer.
- Lasso page: `http://www-stat.stanford.edu/~tibs/lasso.html`