STAT 410 - Linear Regression

Lecture

Meng Li

Department of Statistics

Nov 21, 2017

- By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model.
- However, because it is a discrete process—variables are either retained or discarded—it often exhibits high variance, and so does not reduce the prediction error of the full model.
- Shrinkage methods are more continuous, and do not suffer as much from high variability.

Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.
- The ridge coefficients minimize a penalized residual sum of squares, i.e.,

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\}, \quad (1)$$

  where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
  - Larger value of λ means greater amount of shrinkage.
  - The coefficients are shrunk toward zero.

- An equivalent way to write the ridge problem is

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2,$$

$$\text{subject to} \quad \sum_{j=1}^{k} \beta_j^2 \leq t,$$

  where there is a one-to-one correspondence between the parameters λ and t.
- The size constraint on the coefficients in the ridge regression alleviates the problem of large coefficients (in absolute values) and its high variance, which may be a consequence of multicollinearity.

- Write the objective function in matrix form:

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

- The ridge regression solutions are

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

  - The ridge regression solution is again a linear function of $\mathbf{y}$.
  - The inverse $(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}$ exists even if $\mathbf{X}'\mathbf{X}$ is not of full rank.

- In the case of orthonormal inputs, i.e., $\mathbf{X}'\mathbf{X} = \mathbf{I}$, the ridge estimates are just a scaled version of the least squares estimates, that is,

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \hat{\boldsymbol{\beta}}/(1 + \lambda),$$

where $\hat{\boldsymbol{\beta}}$ are the OLS estimates.

- The lasso is a shrinkage method like ridge, with subtle but important differences. The lasso estimate is defined by

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2,$$

$$\text{subject to} \quad \sum_{j=1}^{k} |\beta_j| \leq t.$$

- We can also write the lasso problem in the equivalent Lagrangian form

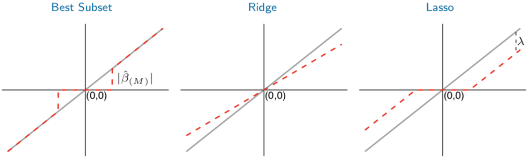
$$\hat{\beta}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\},$$

- The L_2 penalty in the ridge regression $\sum_{j=1}^{k} \beta_j^2$ is replaced by the L_1 penalty $\sum_{j=1}^{k} |\beta_j|$.
- This latter constraint makes the solutions nonlinear in $\mathbf{y}$, and there is no closed form expression as in ridge regression.
- Efficient algorithms such as Least angle regression (LAR) are available for computing the entire path of solutions as λ is varied.
- Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero; this is not obvious.
- Thus the lasso does a kind of continuous subset selection.
- If t is chosen larger than $t_0 = \sum_{j=1}^{k} |\hat{\beta}_j|$ (where $\hat{\beta}_j$ is the OLS estimate), then the lasso estimates are the OLS estimates.

Subset Selection, Ridge Regression and the Lasso

- In the case of an orthonormal input matrix $\mathbf{X}$, the three procedures have explicit solutions.
  - Ridge regression does a proportional shrinkage.
  - Lasso translates each coefficient by a constant factor λ, truncating at zero, i.e., "soft thresholding".
  - Best-subset selection drops all variables with coefficients smaller than the Mth largest, i.e., "hard-thresholding."

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Example: Prostate Cancer

The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`).

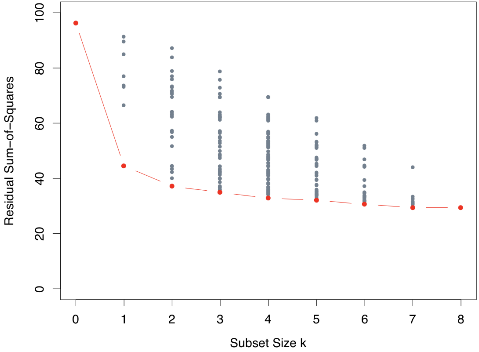
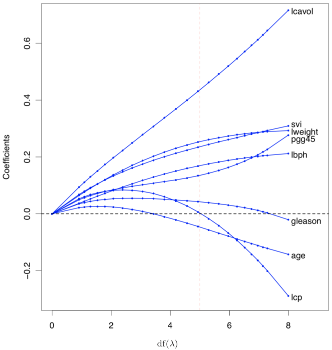
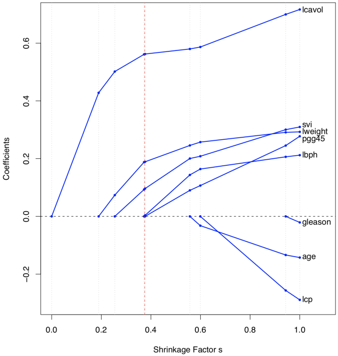


Figure: Profiles of ridge and lasso coefficients. The effective degrees of freedom df(λ) is controlled by λ and defined by tr$(\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}')$. The shrinkage factor s is $t/\sum_{j=1}^{k}|\hat{\beta}_j|$.





Generalization

- We can generalize ridge regression and the lasso:

$$\tilde{\beta} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j|^q \right\},$$

- When $q > 1$, $|\beta_j|^q$ is differentiable at 0, and so does not set the coefficients exactly to zero as in lasso.
- *Elastic-net* penalty uses

$$\lambda \sum_{j=1}^{k} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$$

  as a compromise between ridge and lasso.
- Elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.

A unified framework: Bayesian point of view

- Bayes formula:

$$\pi(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \frac{f(\mathbf{y}|\boldsymbol{\beta},\mathbf{X})\pi(\boldsymbol{\beta})}{\int f(\mathbf{y}|\boldsymbol{\beta},\mathbf{X})\pi(\boldsymbol{\beta})d\boldsymbol{\beta}} \propto f(\mathbf{y}|\boldsymbol{\beta},\mathbf{X})\pi(\boldsymbol{\beta}).$$

- We can view $|\beta_j|^q$ as the log-prior density for β_j.
- The lasso, ridge regression and best subset selection are Bayes estimates with different priors.
- They are derived as posterior modes rather than the posterior mean which is more commonly used in Bayesian literature.
    - Ridge regression is also the posterior mean, but the lasso and best subset selection are not.

The end of the beginning

- Seven pillars of statistical wisdom (Stigler at the JSM 2014)

  Wisdom has built her house;
  She has hewn out her seven pillars.
  - Proverbs 9:1

  1. Aggregation of information.
  2. Diminishing information.
  3. Mathematical quantification of information/uncertainty.
  4. Intercomparisons.
  5. Regression and multivariate analysis.
  6. Design.
  7. Models and residuals.

- Ten Simple Rules for Effective Statistical Practice
  `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004961`