Regularized Regression

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Version: 21/03/2019, 16:40

Regularization

- Which features/variables should we include in a regression function and "how much" of each variable should we include?
- Regularization, also known as shrinkage, is a general approach for model/variable selection and for preventing overfitting.
- The main idea is to introduce additional modeling assumptions or impose constraints on the estimators, usually through a penalty for complexity in the loss function.
- As seen earlier, model/estimator complexity can be measured in various ways, e.g., in regression, number of covariates, magnitude of regression coefficients, smoothness of the regression function.

Regularization

- For linear regression, with the squared error/L_2 loss function, common regularization approaches involve "penalizing" covariates with "large" regression coefficients.
  - Ridge regression: Penalty based on sum of squares (Euclidean/L_2 norm) of regression coefficients.
  - Least absolute shrinkage and selection operator or LASSO: Penalty based on sum of absolute values (L_1 norm) of regression coefficients.
  - Elastic net: Both L_1 and L_2 penalties.
- Regularization techniques may themselves require another layer of model selection, corresponding to the tuning of complexity parameters used to penalize the loss function. Cross-validation may be used for this purpose.

Regularization

- In what follows, we assume we have a learning set $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ that is a random sample of n covariate/outcome pairs from the population of interest.
- Define the design matrix or model matrix $\mathbf{X}_n$ as the $n \times J$ matrix with ith row corresponding to the ith covariate vector X_i, $i = 1, \ldots, n$.
- Define the outcome vector $\mathbf{Y}_n$ as an n-dimensional column vector with ith element corresponding to the ith outcome Y_i, $i = 1, \ldots, n$.

Regularization

- We are interested in fitting linear regression functions of the form

$$E[Y|X] = X^\top \beta = \sum_{j=1}^{J} \beta_j X_j = \beta_1 X_1 + \ldots + \beta_J X_J, \quad (1)$$

where the column vector $\beta = (\beta_j : j = 1, \ldots, J) \in \mathbb{R}^J$ contains the parameters of the model, referred to as regression coefficients.

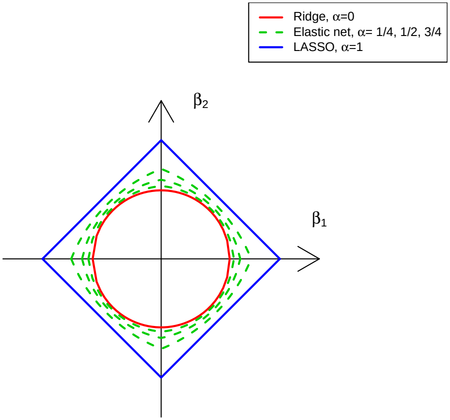

Figure 1: *Elastic net regression.* Constraints for elastic net, $J = 2$: $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \leq \kappa$, $\kappa = 3$.

Rigde Regression

- Ridge regression adds an L_2 penalty for the regression coefficients to the usual squared error loss function, i.e., the estimator of the regression coefficients is defined as

$$\begin{aligned} \hat{\beta}_n^{\text{ridge}} &\equiv \text{argmin}_{\beta \in \mathbb{R}^J} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{J} \beta_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \\ &= (\mathbf{Y}_n - \mathbf{X}_n \beta)^\top (\mathbf{Y}_n - \mathbf{X}_n \beta) + \lambda \beta^\top \beta. \end{aligned} \quad (2)$$

- The shrinkage parameter $\lambda \geq 0$ is a tuning parameter that controls the complexity of an estimator, i.e., the bias-variance trade-off.
- The larger λ, the greater the shrinking of the coefficients toward zero.

Rigde Regression

- One can show, using an argument similar as in the lecture on "Linear Regression", that the ridge regression estimator is

$$\hat{\beta}_n^{\text{ridge}} = (\mathbf{X}_n^\top \mathbf{X}_n + \lambda \mathbf{I}_J)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n. \tag{3}$$

- When $\lambda = 0$, we have the usual linear regression estimator, also known as ordinary least squares (OLS) estimator,

$$\hat{\beta}_n^{\text{OLS}} = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n. \tag{4}$$

- Ridge regression simply adds a positive constant to the diagonal of $\mathbf{X}_n^\top \mathbf{X}_n$, which makes the matrix non-singular, even when the design matrix is not of full rank.
- The ridge estimator is biased, but typically less variable than the ordinary least squares estimator.

Rigde Regression

- As the penalty parameter λ increases, bias tends to increase while variance tends to decrease.

LASSO Regression

- The least absolute shrinkage and selection operator (LASSO) is a shrinkage method similar in spirits to ridge regression, with subtle, yet important differences.
- LASSO regression adds an L_1 penalty for the regression coefficients to the usual squared error loss function, i.e., the estimator of the regression coefficients is defined as

$$\begin{aligned}\hat{\beta}_n^{\text{LASSO}} &\equiv \text{argmin}_{\beta \in \mathbb{R}^J} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{J} \beta_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^{J} |\beta_j| \\ &= (\mathbf{Y}_n - \mathbf{X}_n \beta)^\top (\mathbf{Y}_n - \mathbf{X}_n \beta) + \lambda \|\beta\|_1. \qquad (5)\end{aligned}$$

- The shrinkage parameter $\lambda \geq 0$ is a tuning parameter that controls the complexity of an estimator.

LASSO Regression

- When $\lambda = 0$, we have the usual OLS estimator.
- However, unlike ridge regression, there is no closed-form expression for the LASSO estimator of the regression coefficients.
- Instead, one can rely on a variety of numerical optimization methods to minimize the penalized risk function.
- Also, unlike ridge regression, the LASSO estimator is not linear in the outcome $\mathbf{Y}_n$.
- The LASSO can be used for variable selection. By virtue of the L_1 constraint in Equation (5), making λ sufficiently large causes some of the estimators to be exactly zero.

LASSO Regression

- While there are no closed-form expressions for the bias and variance of the LASSO estimator, bias tends to increase while variance tends to decrease as the amount of shrinkage increases.

Ridge vs. LASSO Regression

- Interpretative ability: Unlike the LASSO, ridge regression does not perform variable selection, in the sense that it does not set regression coefficients exactly to zero unless $\lambda = \infty$, in which case they are all zero.
- Predictive ability: Similar mean squared error.
- Computational complexity: Similar; good algorithms available for the LASSO.

Elastic Net Regression

- The elastic net estimator of the regression coefficients generalizes both ridge and LASSO regression, in that it involves both an L_1 and an L_2 penalty,

$$\hat{\beta}_n^{\text{enet}} \equiv \text{argmin}_{\beta \in \mathbb{R}^J} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{J} \beta_j X_{i,j} \right)^2 + \lambda_1 \sum_{j=1}^{J} |\beta_j| + \lambda_2 \sum_{j=1}^{J} \beta_j^2. \quad (6)$$

- The shrinkage parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters that control the strength of the penalty terms, i.e., the complexity or shrinking of the coefficients towards zero.

Elastic Net Regression

Table 1: *Elastic net regression.* The elastic net covers as special cases the following regression methods.

Method	**Shrinkage parameters**
OLS	$\lambda_1 = \lambda_2 = 0$
Ridge	$\lambda_1 = 0,\ \lambda_2 \geq 0$
LASSO	$\lambda_1 \geq 0,\ \lambda_2 = 0$
Elastic net	$\lambda_1 \geq 0,\ \lambda_2 \geq 0$
$\hat{\beta}_n = 0$	$\lambda_1 = \infty$ or $\lambda_2 = \infty$

Bias-Variance Trade-Off

- As usual, we are faced with a bias-variance trade-off in the choice of the shrinkage parameters λ_1 and λ_2.
- Increasing the amount of shrinkage tends to increase bias and decrease variance.
- By finding the right amount of shrinkage, an increase in bias can be compensated by a decrease in variance, so that risk, here mean squared error (MSE), is reduced overall.
- Cross-validation may be used for tuning the shrinkage parameters.

Pre-Processing the Covariates

- It is usually appropriate to leave the intercept unpenalized, otherwise, the results would depend on the origin chosen for the outcome Y.
- One instead considers centered covariates (mean zero), i.e., a column-centered design matrix $\mathbf{X}_n$. The intercept can then be estimated by the empirical mean of the outcomes $\hat{\beta}_{n,0} = \bar{Y}_n = \sum_i Y_i/n$ and the remaining regression coefficients by regularized regression without intercept.
- The same penalty is imposed on all regression coefficients and their estimators are not invariant to scaling of the covariates. It is therefore common to scale the covariates (to have variance one) prior to performing regularized regression.

Example: Prostate Cancer Dataset

- The prostate specific antigen (PSA) is present in small amounts in the serum of men with healthy prostates, but is often elevated in the presence of prostate cancer or other prostate disorders.
- PSA levels are used in the diagnosis and treatment of prostate cancer.
- We will use the `prostate` dataset to investigate how PSA levels (`lpsa`) relate to the following clinical covariates.
  - `lcavol`: log(cancer volume)
  - `lweight`: log(prostate weight)
  - `age`
  - `lbph`: log(benign prostatic hyperplasia amount)
  - `svi`: seminal vesicle invasion
  - `lcp`: log(capsular penetration)
  - `gleason`: Gleason score

Example: Prostate Cancer Dataset

  - `pgg45`: Percentage Gleason scores 4 or 5.

- The dataset comprises 97 observations in 87 men who were about to undergo a radical prostatectomy.
- The 97 observations are divided into a learning set (LS) of 67 patients and a test set (TS) of 30 patients.
- We will use these data to estimate the regression function of the `lpsa` outcome on the 8 clinical covariates and assess the performance of the estimate in terms of risk for the prediction of PSA levels.

Example: Prostate Cancer Dataset

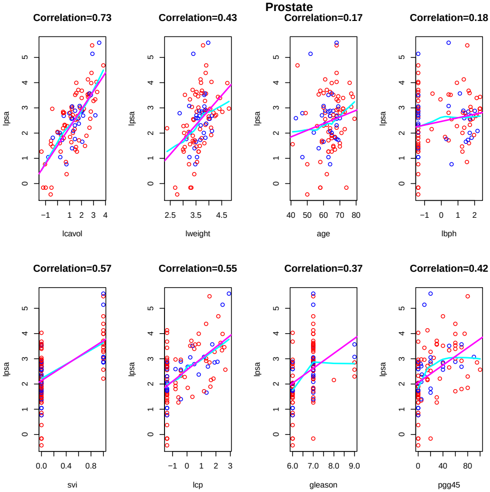

Figure 2: *`prostate` dataset.* Scatterplots of `lpsa` outcome vs. each of 8 covariates (Red: LS, Blue: TS; Magenta: `lm`, Cyan: `lowess`).

Example: Prostate Cancer Dataset

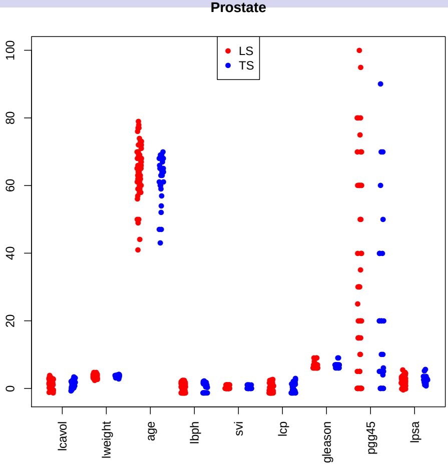

Figure 3: *prostate dataset.* Stripchart of `lpsa` outcome and 8 covariates.

Example: Prostate Cancer Dataset

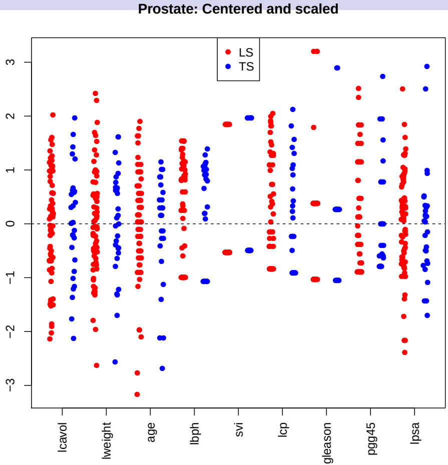

Figure 4: `prostate` *dataset.* Stripchart of `lpsa` outcome and 8 covariates, centered and scaled.

Example: Prostate Cancer Dataset

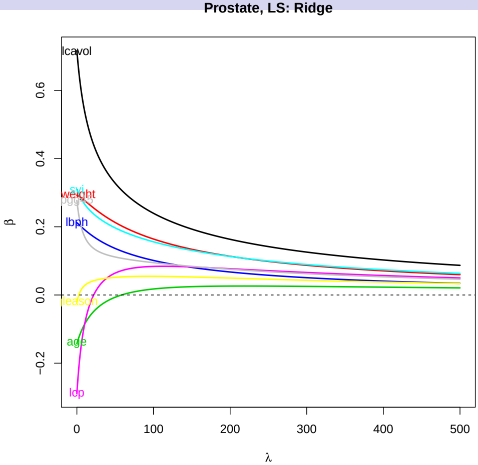

Figure 5: `prostate` *dataset, LS: Ridge.* Estimated regression coefficients vs. shrinkage parameter λ.

Example: Prostate Cancer Dataset

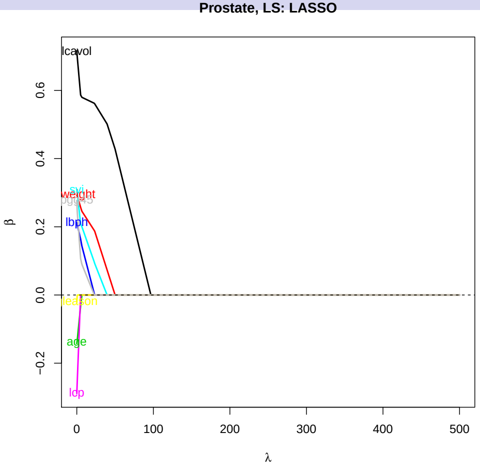

Figure 6: `prostate` *dataset, LS: LASSO.* Estimated regression coefficients vs. shrinkage parameter λ.

Example: Prostate Cancer Dataset

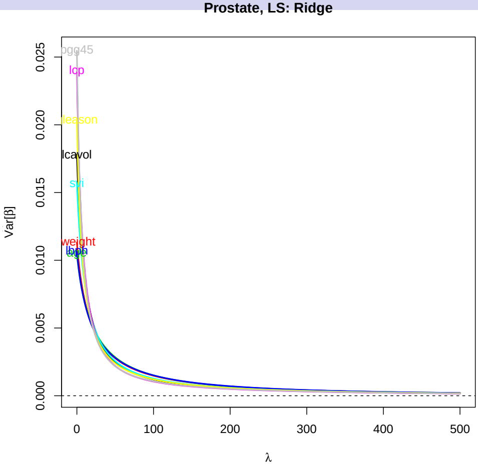

Figure 7: `prostate` *dataset, LS: Ridge.* Estimated variance of estimated regression coefficients vs. shrinkage parameter λ.

Example: Prostate Cancer Dataset

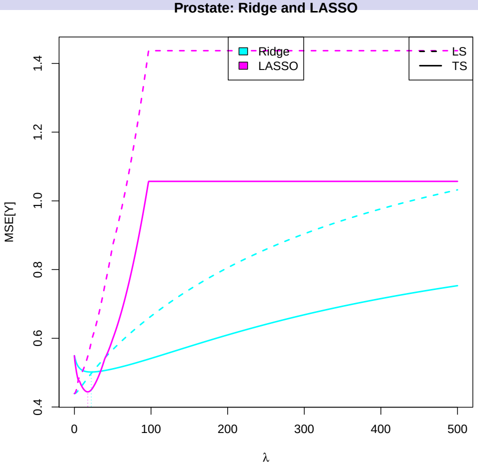

Figure 8: `prostate` *dataset: Ridge and LASSO.* Learning and test set mean squared error vs. shrinkage parameter λ.

Example: Prostate Cancer Dataset

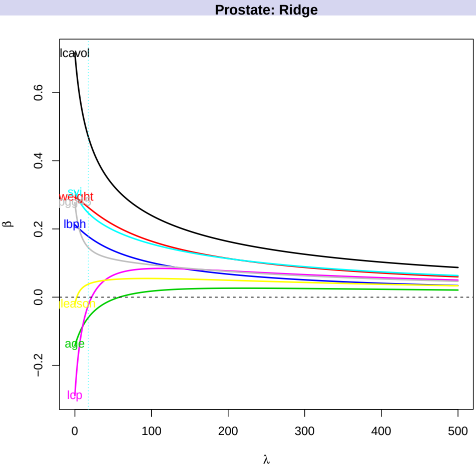

Figure 9: `prostate` *dataset: Ridge.* Estimated regression coefficients vs. shrinkage parameter λ. Dashed line indicates optimal λ based on TS MSE.

Example: Prostate Cancer Dataset

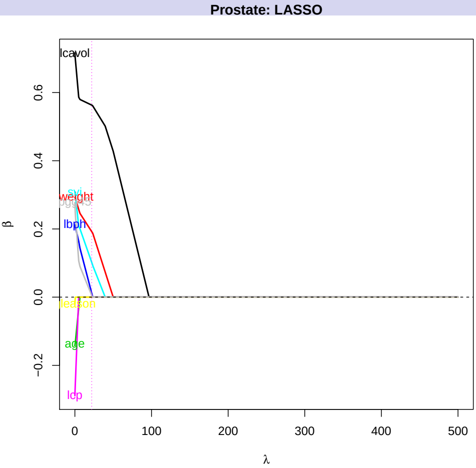

Figure 10: `prostate` *dataset: LASSO.* Estimated regression coefficients vs. shrinkage parameter λ. Dashed line indicates optimal λ based on TS MSE.

Example: Prostate Cancer Dataset

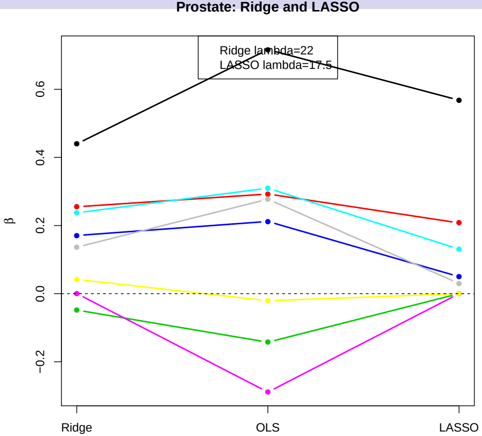

Figure 11: `prostate` *dataset: Ridge and LASSO.* Estimated regression coefficients with optimal λ based on TS MSE.