Regression Shrinkage and Selection via the Lasso[1]

Hsieh Yuan-Hao

2024/09/27

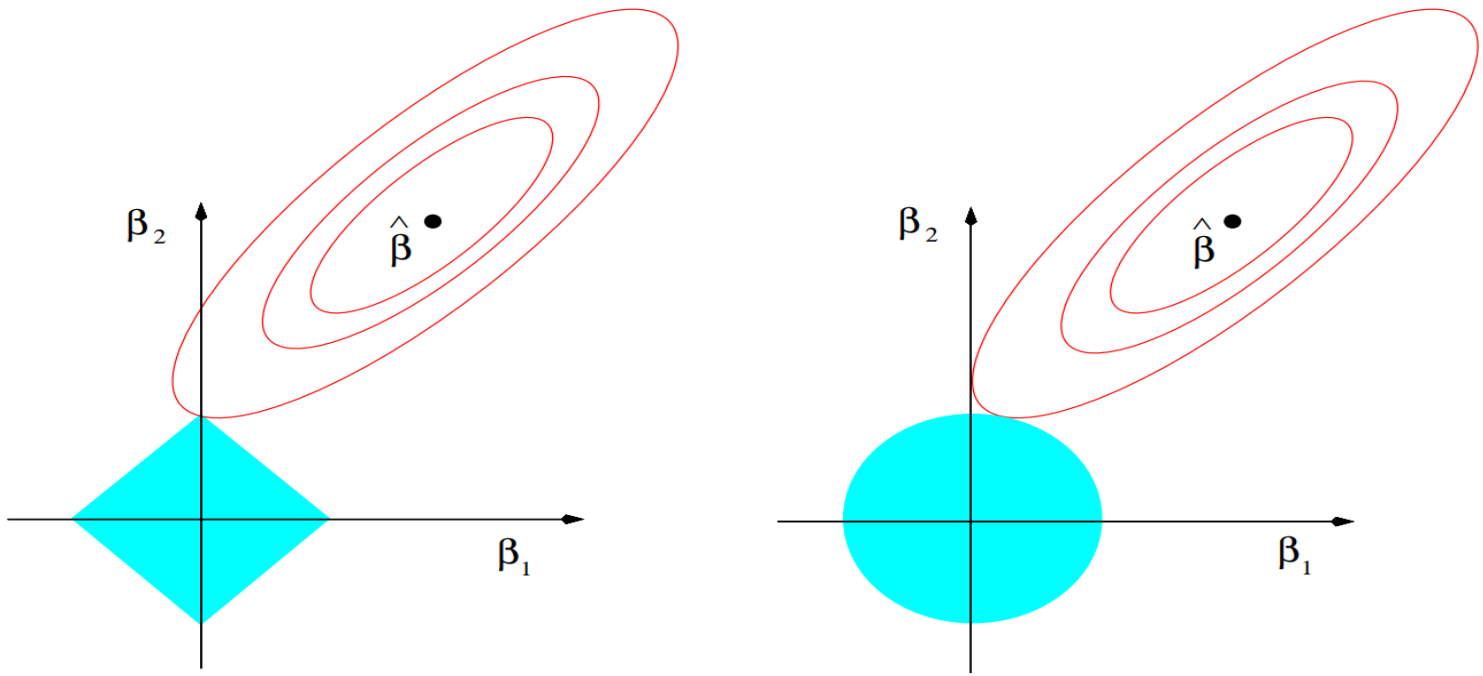
[1]Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

Introduction(1/2)

1. **The Lasso Method**:
   - Proposes a new estimation method for linear models called "lasso" (least absolute shrinkage and selection operator).
2. **Key Features of the Lasso**:
   - Minimizes the residual sum of squares subject to a constraint on the sum of the absolute values of the coefficients.
   - This constraint results in some coefficients being exactly zero, leading to more interpretable models.
3. **Challenges with Ordinary Least Squares (OLS)**:
   - OLS often has low bias but high variance, affecting prediction accuracy.
   - It can be difficult to interpret models with a large number of predictors.

Introduction(2/2)

4. **Comparison with Ridge Regression**:
   - Ridge regression is more stable but does not set coefficients to zero, making the model harder to interpret.
   - Both Lasso and Ridge apply shrinkage, but Ridge uses an L2 penalty, leading to proportional shrinkage of all coefficients.
   - Lasso sets some coefficients to zero, acting as a form of variable selection, while Ridge retains all variables but shrinks their magnitude.



Definition of the Lasso(1/2)

1. **Data Setup**:
   - Observations: (X_i, Y_i), where $X_i = (X_{i1}, \ldots, X_{ip})$ are the predictors and Y_i is the response for $i = 1, 2, \ldots, N$.
   - Standardize X: Each X_{ij} is standardized so that the mean is 0 and the variance is 1.
2. **Lasso Estimate**:
   - The lasso estimate (α, β) minimizes:

$$\min_{\alpha,\beta} \left\{ \sum_{i=1}^{N} \left(Y_i - \alpha - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 \right\}$$

   subject to:

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

   - Tuning Parameter$t$
     1) Controls the amount of shrinkage applied to the coefficients.
     2) When $t < t_0$, where t_0 is the sum of the absolute values of the least squares estimates, some coefficients shrink toward 0.

Definition of the Lasso(2/2)

3. **Optimization**:
   - Solving this problem is a quadratic programming problem with linear inequality constraints
   - Efficient and stable algorithms are available for its computation.
4. **Shrinkage Behavior**:
   - As t decreases, more coefficients are shrunk towards zero, with some becoming exactly zero, depending on the value of t.
5. **Comparison with Non-Negative Garotte**:
   - The motivation for the lasso came from an interesting proposal of Breiman(1993)[2],which is minimizes:

$$\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j c_j\hat{\beta}_j^o x_{ij}\right)^2 \qquad \text{subject to } c_j \geqslant 0, \quad \sum c_j \leqslant t.$$

[2]Breiman, L. (1993). Better subset selection using the non-negative garotte. Tech. Rep., Univ. Calif., Berkeley.

Lasso Solution in Orthonormal Case(1/3)

1. The lasso minimizes the residual sum of squares subject to the ℓ_1-norm constraint on the coefficients β:

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

   Where λ is a tuning parameter controlling the amount of shrinkage.

2. **Orthonormal Design Assumption**
   When $X^T X = I$ (i.e., the design matrix is orthonormal), the problem becomes simpler because the predictors are uncorrelated. In this case, the objective function simplifies for each individual coefficient β_j as:

$$\hat{\beta}_j = \arg\min_{\beta_j} \left((y_j - \beta_j)^2 + \lambda |\beta_j| \right)$$

Lasso Solution in Orthonormal Case(2/3)

3. **Differentiating the Objective Function**:

$$\frac{d}{d\beta_j}\left((y_j - \beta_j)^2 + \lambda|\beta_j|\right) = -2(y_j - \beta_j) + \lambda \cdot \text{sign}(\beta_j)$$

Setting the derivative to zero to find the minimum:

$$-2(y_j - \beta_j) + \lambda \cdot \text{sign}(\beta_j) = 0$$

4. **Solving for β_j**:
   - If $\beta_j > 0$:

$$2(y_j - \beta_j) = \lambda \quad \Rightarrow \quad \hat{\beta}_j = y_j - \frac{\lambda}{2}$$

   - If $\beta_j < 0$:

$$2(y_j - \beta_j) = -\lambda \quad \Rightarrow \quad \hat{\beta}_j = y_j + \frac{\lambda}{2}$$

   - If $|y_j| < \frac{\lambda}{2}$, then the solution is $\hat{\beta}_j = 0$ (because shrinking beyond zero is not possible).

Lasso Solution in Orthonormal Case(3/3)

5. **Soft-Thresholding Function**: This leads to the soft-thresholding rule:

$$\hat{\beta}_j = \text{sign}(y_j) \left(|y_j| - \frac{\lambda}{2} \right)_+$$

where $(z)_+$ denotes the positive part function, meaning $(z)_+ = z$ if $z > 0$, and 0 otherwise.

6. **Final Solution**: The solution for each coefficient in the orthonormal design case is:

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j) \left(|\hat{\beta}_j| - \lambda \right)_+$$

This is the **lasso solution** in the orthonormal case, where the coefficients are shrunk towards zero, and some are set to exactly zero based on the value of the tuning parameter λ.

More on two-predictor case(1/3)

Suppose that p = 2, and assume without loss of generality that the least squares estimates $\hat{\beta}_j$ are both positive.

- **Lasso Objective Function**

$$\min_{\beta_1,\beta_2} \left\{ \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \right\}$$

subject to:

$$|\beta_1| + |\beta_2| \leq t$$

- **The Lagrangian Formulation**

$$L(\beta_1, \beta_2, \lambda) = RSS + \lambda(|\beta_1| + |\beta_2| - t)$$

Since we are assuming that β_1 and β_2 are positive, the absolute values of the coefficients are just the coefficients themselves, so:

$$L(\beta_1, \beta_2, \lambda) = RSS + \lambda(\beta_1 + \beta_2 - t)$$

More on two-predictor case(2/3)

- For β_1:

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n} x_{i1}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2}) + \lambda = 0$$

Simplifying:

$$-2\sum_{i=1}^{n} x_{i1}y_i + 2\beta_1 \sum_{i=1}^{n} x_{i1}^2 + 2\beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \lambda = 0$$

- For β_2:

$$\frac{\partial L}{\partial \beta_2} = -2\sum_{i=1}^{n} x_{i2}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2}) + \lambda = 0$$

Simplifying:

$$-2\sum_{i=1}^{n} x_{i2}y_i + 2\beta_2 \sum_{i=1}^{n} x_{i2}^2 + 2\beta_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \lambda = 0$$

More on two-predictor case(3/3)

- **Solving for β_1 and β_2**: Now, we assume that both coefficients are positive and work on solving the system of equations. The constraint that $\beta_1 + \beta_2 = t$ implies that the sum of the two coefficients is constrained.
- After solving these equations and applying the constraint, the general form for the Lasso estimates of β_1 and β_2 is:

$$\beta_1 = \left(\frac{t}{2} + \frac{\hat{\beta}_1 - \hat{\beta}_2}{2} \right)$$

$$\beta_2 = \left(\frac{t}{2} - \frac{\hat{\beta}_1 - \hat{\beta}_2}{2} \right)$$

Where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the ordinary least squares (OLS) estimates for β_1 and β_2, respectively.

Compare with ridge regression

- The shrinkage behavior of Ridge Regression depends on the correlation between the predictors. Ridge applies proportional shrinkage when predictors are uncorrelated ($\rho = 0$) but differentially shrinks the coefficients when there is a high correlation between the predictors.
- As correlation increases, Ridge tends to shrink the coefficients unequally, while Lasso consistently shrinks them towards zero in a proportional manner.

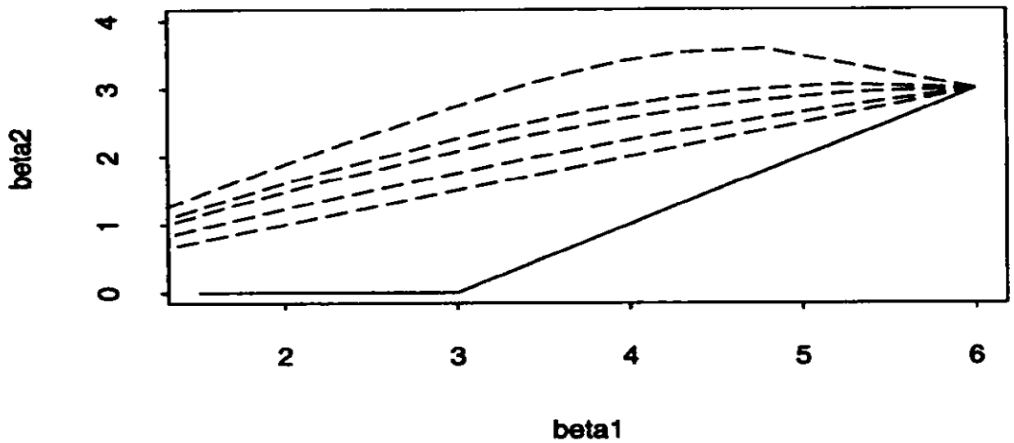


Fig. 4. Lasso (———) and ridge regression (– – – –) for the two-predictor example: the curves show the (β_1, β_2) pairs as the bound on the lasso or ridge parameters is varied; starting with the bottom broken curve and moving upwards, the correlation ρ is 0, 0.23, 0.45, 0.68 and 0.90

Standard Errors(1/3)

1. **Nonlinear and Non-differentiable**:
   - Lasso estimates are challenging to derive standard errors for because of the nonlinearity and non-differentiability introduced by the $|\beta_j|$ penalty.
2. **Bootstrap Approach**:
   - we resample the data and recompute the Lasso estimates β_j over multiple iterations, and compute the variability of the estimates across these samples.
   - Let the Lasso solution for a given bootstrap sample be denoted as $\hat{\beta}_j^{(b)}$. The standard error for each coefficient can then be approximated as:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_j^{(b)} - \bar{\beta}_j\right)^2}$$

   where B is the number of bootstrap samples and $\bar{\beta}_j$ is the average of the $\hat{\beta}_j^{(b)}$ across all bootstrap samples.

Standard Errors(2/3)

3. **Approximate Closed-form Estimate via Ridge Regression**
   - An approximate closed-form formula can be derived by relating Lasso to Ridge regression. This involves approximating the penalty $\sum |\beta_j|$ in Lasso as follows:

$$\sum |\beta_j| \approx \sum \frac{\beta_j}{|\beta_j|}$$

   - At the Lasso solution $\hat{\beta}$, we approximate the solution as:

$$\hat{\beta}^* = (X^T X + \lambda W)^{-1} X^T y$$

   where W is a diagonal matrix with entries $W_{jj} = |\hat{\beta}_j|$, and λ is the regularization parameter. The approximate covariance matrix of the estimates can be computed as:

$$\text{Cov}(\hat{\beta}) \approx \sigma^2 (X^T X + \lambda W)^{-1}$$

   where σ^2 is an estimate of the error variance.

Standard Errors(3/3)

4. **Issue with Zero Coefficients**: A limitation of this approach is that when $\hat{\beta}_j = 0$, the formula will give an estimated variance of 0, i.e., $\text{Var}(\hat{\beta}_j) = 0$ for coefficients that are exactly zero. This occurs because Lasso sets some coefficients to exactly zero for variable selection, which is why the formula gives no uncertainty (variance) for those predictors.
5. **Iterated Ridge Regression Algorithm**:An iterated Ridge regression approach can be used to compute the Lasso estimates, but it is inefficient. The standard error for Ridge is computed as:

$$\text{Cov}(\hat{\beta}) = \sigma^2(X^T X + \lambda I)^{-1}$$

This method can give approximate standard errors for Lasso but may not fully capture the sparsity pattern induced by Lasso.

Prediction Error and Estimation of t (1/3)

1. **Prediction Error Definition**:

$$\text{PE} = E\left[(Y - \hat{f}(X))^2\right]$$

where $\hat{f}(X)$ is the fitted value from the model and Y is the actual response. The goal is to minimize this prediction error.

2. **Cross-Validation (CV) Method**:
   - The data is split into K parts, and the model is trained on $K-1$ parts, while the remaining part is used to estimate the prediction error.
   - This process is repeated for each subset, and the average prediction error is computed. The value of t that minimizes the cross-validation error is selected as the optimal t.
     The general form of CV for Lasso is:

$$\hat{t} = \arg\min_t \frac{1}{K} \sum_{k=1}^{K} \text{PE}_k(t)$$

where $\text{PE}_k(t)$ is the prediction error for the k-th fold.

Prediction Error and Estimation of t (2/3)

3. **Generalized Cross-Validation (GCV)**:
   - **GCV** is a computationally efficient alternative to cross-validation.
   - Instead of explicitly dividing the data into folds, GCV approximates the prediction error based on the residual sum of squares (RSS) and the effective number of parameters ($p(t)$).
   - The GCV formula is:

$$\mathrm{GCV}(t) = \frac{1}{n} \frac{\mathrm{RSS}(t)}{\left(1 - \frac{p(t)}{n}\right)^2}$$

$$\hat{t}_{\mathrm{GCV}} = \arg\min_t \mathrm{GCV}(t)$$

   where $p(t)$ represents the number of effective parameters (a function of t).

Prediction Error and Estimation of t (3/3)

4. Stein's Unbiased Risk Estimate (SURE)

- **SURE** is another method for estimating the risk or prediction error.
- It provides an unbiased estimate of the prediction error under certain conditions, particularly when the errors in the model are normally distributed.
- The SURE formula for Lasso involves penalizing the residual sum of squares based on the number of non-zero coefficients.
- SURE provides an unbiased estimate of RSS, which is:

$$\text{SURE}(t) = \text{RSS}(t) + 2 \cdot \sigma^2 \cdot p(t)$$

where:

- σ^2 is the variance of the errors,
- $p(t)$ is the effective number of parameters in the model.

The optimal value of t is the one that minimizes the SURE estimate:

$$\hat{t}_{\text{SURE}} = \arg\min_t \text{SURE}(t)$$

Lasso as Bayes Estimate(1/2)

1. **Bayesian Interpretation of Lasso**
   - The Lasso constraint $\sum |\beta_j| \leq t$ is equivalent to adding a penalty term proportional to the absolute value of the coefficients in the objective function.
   - This corresponds to assuming **double-exponential (Laplace) priors** for the coefficients β_j.
2. **Double-Exponential Prior**
   - The prior for β_j in the Bayesian interpretation of Lasso is:

$$p(\beta_j) = \begin{cases} \frac{\lambda}{2} \exp(-\lambda \beta_j), & \text{if } \beta_j \geq 0, \\ \frac{\lambda}{2} \exp(\lambda \beta_j), & \text{if } \beta_j < 0. \end{cases}$$

   - The Laplace prior is centered at zero and heavily penalizes large coefficients, leading to the shrinkage behavior seen in Lasso.

Lasso as Bayes Estimate(2/2)

3. **Comparison with Ridge Regression**:
   a. Prior Distribution:
      - Lasso: The coefficients β_j have a **Laplace (double-exponential)** prior:

$$p(\beta_j) = \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right)$$

      - Ridge: The coefficients β_j have a **Gaussian (Normal)** prior:

$$p(\beta_j) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda\beta_j^2}{2}\right)$$

   b. Shrinkage Behavior
      - Lasso shrinks some coefficients to exactly zero, leading to sparse solutions and variable selection.
      - Ridge regression shrinks all coefficients, but none of them are exactly zero. The coefficients are continuously shrunk towards zero, but Ridge does not perform variable selection.
   c. Penalization
      - Lasso penalizes the absolute values of the coefficients: $\sum|\beta_j|$.
      - Ridge penalizes the square of the coefficients: $\sum\beta_j^2$.

Example - Prostate Cancer Data (to be continued...)

- This section illustrates the application of Lasso using the Prostate Cancer dataset[3], a well-known dataset used to study the relationship between several clinical measures and the level of prostate-specific antigen (PSA), a marker used to detect prostate cancer.

lpsa	log prostate specific antigen
lcavol	log cancer volume
lweight	log prostate weight
age	age
lbph	log of benign prostatic hyperplasia amount
svi	seminal vesicle invasion
lcp	log of capsular penetration
gleason	Gleason score
pgg45	percent of Gleason scores 4 or 5

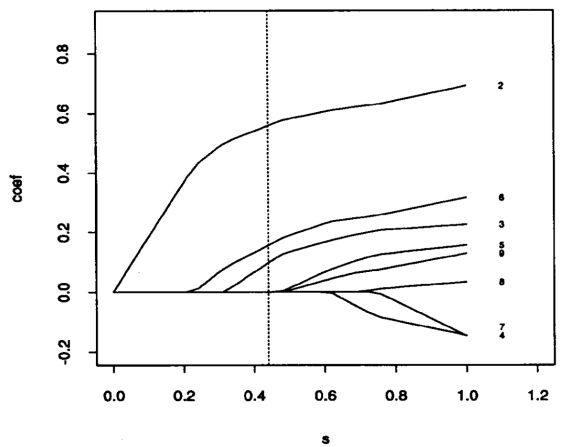


Fig. 5. Lasso shrinkage of coefficients in the prostate cancer example: each curve represents a coefficient (labelled on the right) as a function of the (scaled) lasso parameter $s = t/\Sigma|\hat{\beta}_j^0|$ (the intercept is not plotted); the broken line represents the model for $\hat{s} = 0.44$, selected by generalized cross-validation

[3]Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., Yang, N. (1989).