# Regularization and Variable Selection

Wei Li

Syracuse University

Spring 2024

# OVERVIEW

Motivation

Model Selection

Lasso (least absolute shrinkage and selection operator)

Ridge Regression

Computation of LASSO

Consistency of LASSO

Beyond Lasso and Remarks

# Motivation

# Predictive Accuracy

Suppose $Y_i = \beta^\top X_i + \epsilon_i$, $X_i$ a p-vector, where $E\left[\epsilon_i \mid X_i\right] = 0$ and $\operatorname{Var}\left(\epsilon_i \mid X_i\right) = \sigma^2$.

- For a truly linear model, the bias of the OLS prediction is exactly 0, and the variance is $p \cdot \sigma^2/n$
- When n $\gg p$, OLS tends to have low variance, and hence perform well on test observations.
- When $n < p$, then there is no longer OLS estimate, and the variance of these estimates are infinite. May be able to increase bias slightly, but decrease variance substantially via
  - regularization/shrinkage
  - features selection
  - dimension reduction

The **key message**: It is possible to trade a little bias with the large reduction in variance, thus achieving higher prediction accuracy.

# Model Interpretability

- When we have a large number of variables $X$ in the model there will generally be many that have little or no effect on $Y$
- Leaving these variables in the model makes it harder to see the "big picture", i.e., the effect of the "important variables"
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables. This can be accomplished via variable selection, or feature selection.

# Model Selection

# Best Subset Selection

1. For each $k \in \{0, 1, \ldots, p\}$, find the subset of size $k$ that gives smallest residual sum of squares
   - fit all $\binom{p}{k}$ models that contains $k$ predictors
   - pick the best subset (model) with smallest RSS or $R^2$.
2. Then among the $p + 1$ chosen models, pick a single best model using some criterion discussed below.

# Sequential Selection

- **Forward Stepwise Selection**: Begins with the model containing no predictor, and then adds one predictor at a time that improves the model the most (say, by $R^2$, or RSS ) until no further improvement is possible.
- **Backward Stepwise Selection**: Begins with the model containing all predictors, and then deleting one predictor at a time that improves the model the most (say, by $R^2$, or RSS ) until no further improvement is possible.

Note that both procedures produce a sequence of models of different size $k$ (i.e., different number of predictors $k$).

Note that the larger $k$ (model size), the smaller RSS, or higher $R^2$. To choose the best model size $k^*$, one can use a number of different criteria:

- adjusted-$R^2$
- prediction error on the test set
- Mallow's $C_p$, AIC, BIC
- cross validation
    - generalized cross validation (GCV) (later)
- bootstrap (later)

# Cross Validation

Use cross validation to choose the model size:

Outer Loop (CV): split the dataset into training and validation sets

- For each fold:
    - model selection: on the training set of each CV split, run best-subset selection, forward stepwise (backwawrd stepwise) selection, yielding fitted models of different sizes
    - validation: evaluate the performance on the validation set of the CV split for all models of different sizes.
- Average the validation performance over all folds
- Choose the best model size $\hat{k}$ that has the smallest averaged validation error

Run the selection & estimation on the whole dataset with the model size $\hat{k}$.

# Adjustment for traininig errors

Suppose the data are $(\boldsymbol{x}_i, y_i)\,, i = 1, \cdots, n$. A fitted linear regression model is $\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\beta}}^\top \mathbf{x}$.

- The degree of freedom (df) of $\hat{\boldsymbol{\beta}}$ as the number of nonzero elements in $\boldsymbol{\beta}$ (model size, say $k$), including the intercept
- Adjusted $R^2 = 1 - \frac{\text{RSS}\,/(n-k)}{\text{TSS}\,/(n-1)}$
- The residual sum of squares $\text{RSS} = \sum_{i=1}^n \left(y_i - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i\right)^2$

$$\begin{aligned}\text{AIC}(k) &= n\log(\text{RSS}/n) + 2\cdot k\\ \text{BIC}(k) &= n\log(\text{RSS}/n) + \log(n)\cdot k\end{aligned}$$

- $$C_p(k) = \frac{1}{n}\left(RSS + 2k\cdot\tilde{\sigma}^2\right),$$

  $\tilde{\sigma}^2$ is an estimate of $\sigma^2$ using a full model.

*AIC, BIC and $C_p$ estimate “in-sample test error”. Details in future lessons.

# Lasso (least absolute shrinkage and selection operator)

# Lasso (least absolute shrinkage and selection operator)

A generalized regularization framework:

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) + \lambda J(\boldsymbol{\beta}), \lambda \geq 0,$$

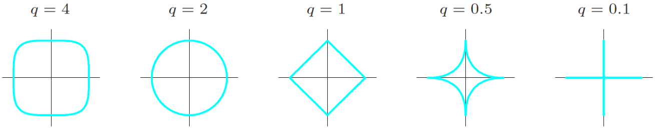
$L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$ is the loss function.

- Squared error loss, $L = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$
- Others: negative log-likelihood, deviance, cross entropy, 0-1 loss, hinge loss function (SVM), exponential loss (AdaBoost),...

$J(\beta)$ is the penalty function.

$$J_q(|\beta|) = \|\beta\|_q^q = \sum_{j=1}^{p} |\beta_j|^q, \quad q \geq 0$$

- $J_0(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} 1\,(\beta_j \neq 0)$ (Best subset; Donoho and Johnstone 1988.)
- $J_1(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ (LASSO; Tibshirani 1996)
- $J_2(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ (Ridge; Hoerl and Kennard, 1970)
- $J_\infty(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_\infty = \max_j \mid \beta_j \mid$ (Supnorm penalty; Zhang et al. 2008)



ESL: Fig 3.12 contours of constant value of $\sum_j |\beta_j|^q$

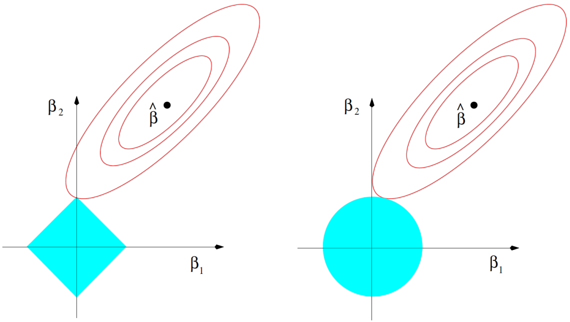
For the LS problem, $\lambda > 0$,

$$\begin{array}{lll} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} & \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_0 & \text{(Best subset selection)} \\ \min_{\boldsymbol{\beta}\in\mathbb{R}^p} & \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1 & \text{(Lasso regression)} \\ \min_{\boldsymbol{\beta}\in\mathbb{R}^p} & \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2 & \text{(Ridge regression)} \end{array}$$

Alternatively, for some $t > 0$:

$$\begin{array}{l} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_0 \le t \text{ (Best subset selection)} \\ \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 \le t \text{ (Lasso regression)} \\ \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_2^2 \le t \text{ (Ridge regression)} \end{array}$$

*not exact equivalence for best subset method.



- Solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$,
- Red ellipses are the contours of the least squares error function.

ESL: Fig 3.11

# Lasso

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- For estimation, we center predictors and response, and standardize predictors (important).
  - For prediction, we scale back the coefficients to work with un-transformed predictors.

- $\lambda \geq 0$ is a tuning parameter
- $\lambda$ controls the amount of shrinkage; the larger $\lambda$, the greater amount of shrinkage
- What happens if $\lambda \to 0$? (no penalty); $\lambda \to \infty$?
- No explicit solution in general.

The equivalent optimization problem is

$$\hat{\beta}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

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

| | small | large |
|---|---|---|
| $t$ | less complex/underfitting | more complex/over-fitting |
| $\lambda$ | more complex/over-fitting | less complex/underfitting |

- a kind of continuous subset selection
- let $t_0 = \sum_{j=1}^{p} \left| \hat{\beta}_j^{ols} \right|$
    - standardize $t$ by $s = t/t_0$ (**shrinkage factor**) in $[0, 1]$

# Special Case: orthonormal design matrix

Assume that $\mathbf{X}^\top\mathbf{X} = I$:

$$\min_{\beta}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$$

The solution to the Lasso problem is

$$\begin{aligned}\hat{\beta}_j^{\text{lasso}} &= \text{sign}\left(\hat{\beta}_j^{ols}\right)\left(\left|\hat{\beta}_j^{ols}\right| - \frac{\lambda}{2}\right)_+ \\ &= \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{ols} > \frac{\lambda}{2} \\ 0 & \text{if } \left|\hat{\beta}_j^{ols}\right| \leq \frac{\lambda}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2} & \text{if } \beta_j^{ols} < -\frac{\lambda}{2} \end{cases}\end{aligned}$$

“soft-thresholding”

- shrinks big coefficients by a constant $\frac{\lambda}{2}$ towards zero.
- truncates small coefficients to zero exactly.

# General Case

$$\min_{\beta}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j|$$

The solution $\hat{\boldsymbol{\beta}}_\lambda$ satisfies for some $\tilde{\boldsymbol{\gamma}}\in\mathbb{R}^p$ that

$$\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda=\mathbf{X}^\top\mathbf{y}-\lambda\tilde{\boldsymbol{\gamma}}/2$$

where

$$\tilde{\boldsymbol{\gamma}}_j=\begin{cases}\operatorname{sign}(\hat{\boldsymbol{\beta}}_\lambda)_j & \text{if } (\hat{\boldsymbol{\beta}}_\lambda)_j\neq 0\\ \in[-1,1] & \text{if } (\hat{\boldsymbol{\beta}}_\lambda)_j=0\end{cases}.$$

# Choose $\lambda$ by CV

The tuning parameter $t$ or $\lambda$ should be adaptively chosen to minimize the PE (prediction error). Recall CV:

1. Randomly split the data into $K$ roughly equal parts.
2. For each $k = 1, \ldots, K$
   - leave the $k$ th portion out, and fit the model using the other $K - 1$ parts. Denote the solution by $\widehat{\boldsymbol{\beta}}^{-(k)}$
   - calculate prediction errors of $\widehat{\boldsymbol{\beta}}^{-(k)}$ on the left-out $k$ th portion
3. Average the prediction errors.
4. Choose the $k$ that minimizes the average prediction error.

Choose a sequence of $\lambda$ values, say $\{\lambda_1, \ldots, \lambda_m\}$:

- For each fold $k$:
  - For each $\lambda$, fit the LASSO on the training data of fold $k$, and denote the solution by $\hat{\boldsymbol{\beta}}_\lambda^{-(k)}$.
- Compute the $CV(\lambda)$ curve as

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda^{-(\kappa(i))} \right)^2$$

  which provides an estimate of the test error curve.
- Find the best parameter $\hat{\lambda}$ which minimizes $CV(\lambda)$.

$$\hat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots \lambda_m\}}{\operatorname{argmin}} \mathrm{CV}(\lambda)$$

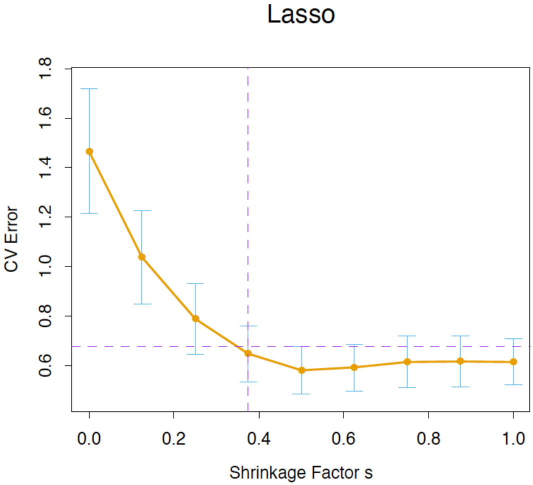
- Fit the final LASSO model with $\hat{\lambda}$ on the whole dataset. The final solution is $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}^{\text{lasso}}$

## One-standard error rule

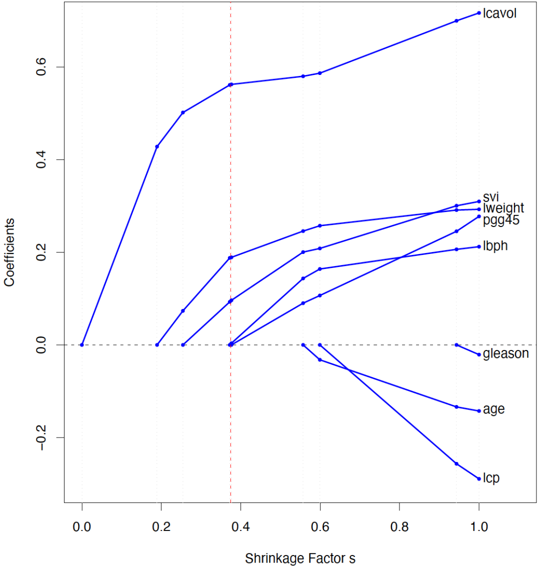
- choose the most parsimonious model with error no more than one standard error above the best error.
- used often in model selection (for a smaller model size).

$$\text{choose } \arg\max \left\{ \lambda \in \{\lambda_1, \dots \lambda_m\} : \mathrm{CV}(\lambda) \leq \mathrm{CV}(\hat{\lambda}) + \mathrm{SE}\left(\mathrm{CV}(\hat{\lambda})\right) \right\}$$

# Prostate Cancer Data Example



ESL Figure 3.7



Key feature: the solution profiles $\lambda \rightarrow \hat{\beta}_\lambda$ are piece-wise linear.

ESL Figure 3.10

# Ridge Regression

# Ridge Regression

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

An equivalent problem

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$
$$\text{subject to } \textstyle\sum_{j=1}^{p} \beta_j^2 \leq t$$

*Assume that the predictors are centered and standardized, and the response centered.

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta$$

The ridge regression solutions is

$$\hat{\beta}^{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

**Theorem**: For any design matrix $\mathbf{X}$, the quantity $\mathbf{X}^\top \mathbf{X} + \lambda I_p$ is always invertible for $\lambda > 0$; thus, there is always a unique solution $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ (Remember, $\mathbf{X}^\top \mathbf{X}$ must be PSD, but not necessarily PD).

In addition, there always exists a parameter $\lambda^* > 0$ such that the MSE of $\hat{\beta}^{\text{ridge}}$ is strictly smaller than the MSE of $\hat{\beta}^{\text{ols}}$ .

Effect of $\lambda$:

- As $\lambda \to 0, \widehat{\boldsymbol{\beta}}^{\text{ridge}} \to \widehat{\boldsymbol{\beta}}^{\text{ols}}$
- As $\lambda \to \infty, \widehat{\boldsymbol{\beta}}^{\text{ridge}} \to \mathbf{0}$

With **orthonormal design matrix**:

$$\hat{\beta}_j^{\text{ridge}} = \frac{\hat{\beta}_j^{\text{ols}}}{1+\lambda}$$

This illustrates the essential “shrinkage” feature of ridge regression

- Applying the ridge regression penalty has the effect of shrinking the estimates toward zero
- The penalty introducing bias but reducing the variance of the estimator
- Ridge estimator does not threshold, since the shrinkage is smooth (proportional to the original coefficient).

# General case: shrinkage

The SVD of $\mathbf{X}$ $(n \times p)$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

- Here $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthonormal and orthogonal matrices, with the columns of U spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space.
- $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_j = 0$, $\mathbf{X}$ is singular.

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \sum_{j=1}^{p} \frac{d_j}{d_j^2 + \lambda} \mathbf{v}_j \mathbf{u}_j^\top \mathbf{y}, \qquad \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ridge}} = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^\top \mathbf{y}$$

$$\text{if all } d_j > 0, \quad \hat{\boldsymbol{\beta}}^{\text{ols}} = \sum_{j=1}^{p} \frac{1}{d_j} \mathbf{v}_j \mathbf{u}_j^\top \mathbf{y}, \qquad \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \sum_{j=1}^{p} \mathbf{u}_j \mathbf{u}_j^\top \mathbf{y}$$

**Principal component analysis** (PCA, Pearson 1901) is a statistical procedure that

- converts a set of observations of correlated variables into a set of linearly uncorrelated variables (called principal components)
- finds directions with maximum variability

The transformed data matrix is given as $\mathbf{Z} = \mathbf{X}V$ where $V$ is obtained through SVD of $\mathbf{X}$. Columns of $V$ give the **principal component (PC) directions**.

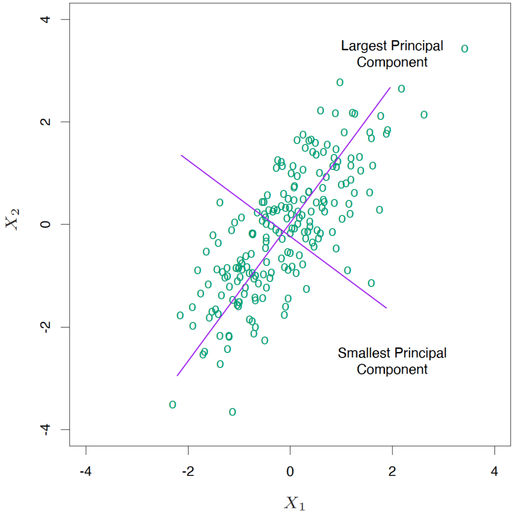
- $\mathbf{z}_1 = \mathbf{X}v_1$: first principal component (PC) of $\mathbf{X}$.
- $\mathbf{z}_{i1}$:$i$-th score of the first (PC).
  - $\mathbf{z}_{i1} = v_1^\top x_i$: projection of $i$-th example on the first PC direction.

The **first principal component direction** $v_1$ has the property that $\mathbf{z}_1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$:

$$\operatorname{Var}(\mathbf{z}_1) = \operatorname{Var}(\mathbf{X}v_1) = \frac{d_1^2}{n}$$

Subsequent principal components $\mathbf{z}_j$ have maximum variance $d_j^2/n$, subject to being orthogonal to the earlier ones. The last principal component has minimum variance.

- the small singular values $d_j$ correspond to directions in the column space of $\mathbf{X}$ having small variance
- ridge regression most effectively shrinks along the directions for small $d_j$



ESL Figure 3.9

# Degrees of Freedom

Assuming $Y|X$ has constant variance $\sigma^2$.

For linear smoother $\hat{\mathbf{y}} = \mathbf{W}\mathbf{y}$, $\mathbf{W}$ “smoothing matrix” depends on

- the training data $\mathbf{X}$
- possibly some tuning parameter, but not $\mathbf{y}$

The “effective degree of freedom” (details in Sec 7.6) is

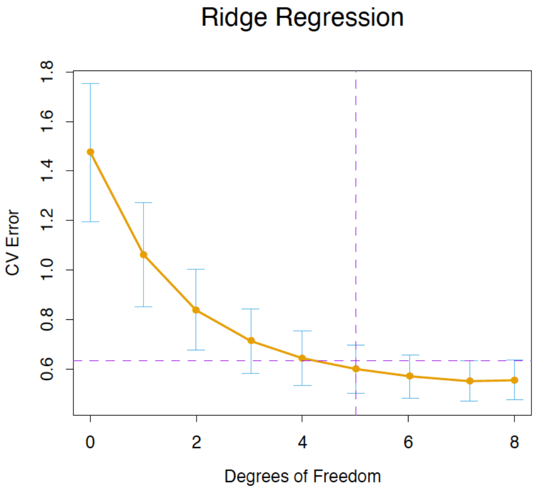
$$\mathrm{df}(\lambda) = \mathrm{trace}(\mathbf{W}).$$

For the ridge estimator,

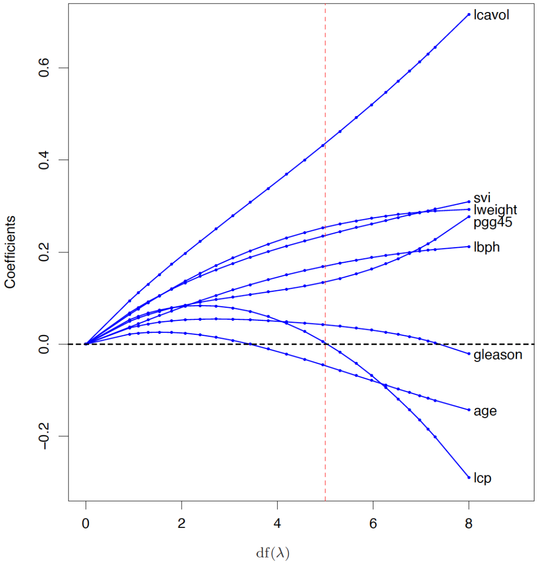
$$\begin{aligned}\mathrm{df}(\lambda) = \mathrm{df}\left(\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}\right) &= \mathrm{trace}\left\{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda \boldsymbol{I}_d\right)^{-1}\mathbf{X}^\top\right\} \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}\end{aligned}$$

Using a more general definition of “effective d.f.”, for the lasso estimator, $\mathrm{df}(\lambda) = \mathrm{df}\left(\hat{y}_\lambda^{\text{lasso}}\right) = \mathrm{E}(\#\text{ of nonzero components in } \hat{\beta}_\lambda^{\text{lasso}})$

# Prostate Cancer Data Example



ESL Figure 3.7



ESL Figure 3.8

# A comparison between ridge regression and OLS

- If the design matrix $\mathbf{X}$ is not full rank, then $\mathbf{X}^\top \mathbf{X}$ is not invertible. For any design matrix $\mathbf{X}$, the quantity $\mathbf{X}^\top \mathbf{X} + \lambda I_p$ is always invertible; thus, there is always a unique solution $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$.
- There always exists a $\lambda$ such that the MSE of $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ is less than the MSE of $\widehat{\boldsymbol{\beta}}^{\text{ols}}$.
- Recall that if the true model is approximately linear, the OLS estimates generally have low bias but can still be highly variable. The penalty term makes the ridge regression estimates biased but can also substantially reduce variance.

# A comparison between ridge regression and the lasso

Ridge regression:

- is a continuous shrinkage method; achieves better prediction performance than OLS through a biase-variance trade-off.
- cannot produce a parsimonious model, for it always keeps all the predictors in the model.

The lasso regression:

- does both continuous shrinkage and automatic variable selection. simultaneously. The lasso not only set coefficients to zero exactly, but it also shrinks the nonzero coefficients.

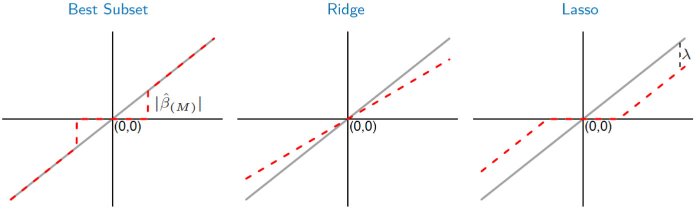
# A comparison for orthonormal design: $\mathbf{X}^\top\mathbf{X} = I$

- Best subset (of size $M$) : $\hat{\beta}_j^{ols}$ if rank $\left(\left|\hat{\beta}_j^{ols}\right|\right) \leq M$ keeps the largest coefficients; "hard-threshholding"
- Ridge regression: $\hat{\beta}_j^{\text{ols}}/(1+\lambda)$: does a proportional shrinkage.
- Lasso:

$$\begin{aligned}\hat{\beta}_j^{\text{lasso}} &= \text{sign}\left(\hat{\beta}_j^{ols}\right)\left(\left|\hat{\beta}_j^{ols}\right| - \frac{\lambda}{2}\right)_+ \\ &= \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{ols} > \frac{\lambda}{2} \\ 0 & \text{if } \left|\hat{\beta}_j^{ols}\right| \leq \frac{\lambda}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2} & \text{if } \beta_j^{ols} < -\frac{\lambda}{2} \end{cases}\end{aligned}$$

"soft-threshholding"

| 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 | $\text{sign}(\hat\beta_j)(\lvert\hat\beta_j\rvert-\lambda)_+$ |



ESL Table 3.4: note that the lasso expression uses the $RSS/2$ criterion.

# Computation of LASSO

# Computation of LASSO

- LARS solution path (Efron et al. 2004)
  - the most efficient algorithm for LASSO
  - designed for standard linear models
  - R package "lars"
  - Python "sklearn.linear_model.Lars"
- Coordinate Descent Algorithm (CDA; Friedman et al. 2010)
  - designed for GLM
  - suitable for ultra-high dimensional problems
  - R package "glmnet"
  - Python "sklearn.linear_model.ElasticNet"

# Coordinate descent for lasso

$f : \mathbb{R}^n \to \mathbb{R}$ convex, differentiable, if we are at a point $x$ such that $f(x)$ is minimized along each coordinate axis, then we found a global minimizer.

Coordinate descent: start with some initial guess $x^{(0)}$, and repeat

$$\begin{aligned}
&x_1^{(k)} \in \underset{x_1}{\operatorname{argmin}} f\left(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \ldots x_n^{(k-1)}\right) \\
&x_2^{(k)} \in \underset{x_2}{\operatorname{argmin}} f\left(x_1^{(k)}, x_2, x_3^{(k-1)}, \ldots x_n^{(k-1)}\right) \\
&x_3^{(k)} \in \underset{x_3}{\operatorname{argmin}} f\left(x_1^{(k)}, x_2^{(k)}, x_3, \ldots x_n^{(k-1)}\right) \\
&\ldots \\
&x_n^{(k)} \in \underset{x_n}{\operatorname{argmin}} f\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots x_n\right)
\end{aligned}$$

for $k = 1, 2, 3, \ldots$.

$$\partial_{\beta_j} RSS^{\text{lasso}}(\boldsymbol{\beta}) = \partial_{\beta_j} RSS^{OLS}(\boldsymbol{\beta}) + \partial_{\beta_j} \lambda \|\boldsymbol{\beta}\|_1$$

Using subdifferential,

$$\beta_j = \beta_j(\rho_j) = \frac{1}{z_j}\text{soft}(\rho_j; \lambda),$$

where $\text{soft}(\rho_j; \lambda) := \text{sign}(\rho_j)(|\rho_j| - \lambda)_+$.

**Algorithm** (Coordinate descent for lasso):

```
Repeat
    for $j = 1, \ldots, d$ do
        $z_j = 2\sum_{i=1}^{n} (x_{ij})^2$
        $\rho_j = 2\sum_{i=1}^{n} x_{ij}[y_i - \beta^\top x_i + \beta_j x_{ij}]$
        $\beta_j = \text{soft}(\frac{\rho_j}{z_j}; \frac{\lambda}{z_j})$
until convergence
```

# Consistency of LASSO

# Consistency of LASSO

It does not always have the oracle property,that is, it does not always perform as well in terms of variable selection as if the true underlying model has been given (Fan and Li, 2001).

Certain conditions are needed for the lasso to have the oracle property (Fan and Li, 2001; Zou, 2006).

True model: $y_i = \beta_{00} + \sum_{j=1}^{p} x_{ij}\beta_{j0} + \epsilon_i$.

important index set: $\mathcal{A}_0 = \{j : \beta_{j0} \neq 0, j = 1, \cdots, p\}$.

unimportant index set: $\mathcal{A}_0^c = \{j : \beta_{j0} = 0, j = 1, \cdots, p\}$.

An oracle performs as if the true model were known:

- selection consistency

$$P\left(\hat{\beta}_j \neq 0 \text{ for } j \in \mathcal{A}_0; \quad \hat{\beta}_j = 0 \text{ for } j \in \mathcal{A}_0^c\right) \to 1$$

- estimation consistency:

$$\sqrt{n}\left(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0}\right) \to_d N\left(0, \Sigma_I\right)$$

$\boldsymbol{\beta}_{\mathcal{A}_0} = \{\beta_j, j \in \mathcal{A}_0\}$ and $\Sigma_l$ is the covariance matrix if knowing the true model.

Knight and Fu (2000)

- Estimation Consistency: The LASSO solution is of estimation consistency for fixed $p$. In other words,

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda_n) \to_p \boldsymbol{\beta}, \text{ as } \lambda_n = o(n)$$

  It is root- $n$ consistent and asymptotically normal.
- Model Selection Property: For $\lambda_n \propto n^{\frac{1}{2}}$, as $n \to \infty$, there is a non-vanishing positive probability for lasso to select the true model.

Zhao and Yu (2006):

Under the Irrepresentable Condition (IC), the LASSO is model selection consistent in both fixed and large $p$ settings.

$$\max_{j \in \mathcal{A}_0^c} \left\| \left( \mathbf{X}_{\mathcal{A}_0}^\top \mathbf{X}_{\mathcal{A}_0} \right)^{-1} \mathbf{X}_{\mathcal{A}_0}^\top \mathrm{x}_j \right\|_1 \leq (1 - \epsilon) \text{ for some } \epsilon \in (0, 1]$$

The least squares coefficients for the columns of $\mathbf{X}_{\mathcal{A}_0^c}$ on $\mathbf{X}_{\mathcal{A}_0}$ should not be too large.

# Beyond Lasso and Remarks

# Elastic net (Zhou and Hastie 2005)

$$J_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left( \lambda \left| \beta_j \right| + (1-\lambda)\beta_j^2 \right), \lambda \in [0,1]$$

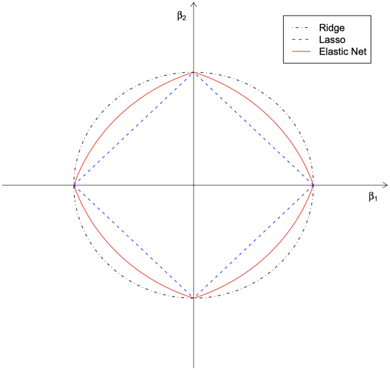


Figure: Elastic net penalty

$$L\left(\lambda_1, \lambda_2, \boldsymbol{\beta}\right) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$
$$\|\boldsymbol{\beta}\|_1 = \textstyle\sum_{j=1}^{p} \left|\beta_j\right|, \quad \text{and } \|\boldsymbol{\beta}\|_2 = \left(\textstyle\sum_{j=1}^{p} \beta_j^2\right)$$

The elastic net penalty function

- has singularity at the vertex (necessary for sparsity)
- has strict convex edges (encouraging grouping)

Then the naive Elastic Net estimator:

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} L\left(\lambda_1, \lambda_2, \boldsymbol{\beta}\right)$$

.

Properties: the elastic net solution

- simultaneously does automatic variable selection and continuous shrinkage
- can select groups of correlated variables, retaining ‘all the big fish’.

A correction to the double shrinkage is to use $(1 + \lambda_2)\hat{\boldsymbol{\beta}}_{\text{EN}}$ as the EN estimator.

# Group Lasso (Yuan and Lin 2006)

- In a regression model, a multi-level categorical predictor is usually represented by a group of dummy variables.
- In an additive model, a continuous predictor may be represented by a group of basis functions to incorporate nonlinear relationship.

Suppose the $p$ predictors are divided into L groups, with $p_\ell$ in group $\ell$.

$$\widehat{\boldsymbol{\beta}}^{\text{glasso}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{L} \mathbf{x}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{L} \sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|$$

where

- $\lambda \geq 0$ is a tuning parameter.
- $\left\| \boldsymbol{\beta}_j \right\| = \sqrt{\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j}$ is the $L_2$ norm.
- The group lasso penalty encourages sparsity at the factor level.
- When $p_1 = \cdots = p_L = 1$, the group lasso reduces to the lasso.

# Sparse-Group Lasso (Simon et al 2013)

The group lasso does not, however, yield sparsity within a group. A modified version is

$$\hat{\boldsymbol{\beta}}^{\text{sglasso}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{L} \mathbf{x}_j \boldsymbol{\beta}_j \right\|^2 + \lambda_1 \sum_{j=1}^{L} \sqrt{p_j}\, \|\boldsymbol{\beta}_j\| + \lambda_2 \|\boldsymbol{\beta}\|_1$$

# Bayesian interpretation of Ridge, Lasso

Model:

$$\mathbf{y} \mid \mathbf{X}, \beta \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right), \quad \boldsymbol{\beta} \sim \mathcal{N}(0, \tau^2\mathbf{I})$$

Posterior (multivariate normal regression semi-conjugate prior):

$$\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\mu_{\boldsymbol{\beta}|\mathbf{X},\mathbf{y}}, \Sigma_{\boldsymbol{\beta}|\mathbf{X},\mathbf{y}}\right)$$

where

$$\mu_{\boldsymbol{\beta}|\mathbf{X},\mathbf{y}} = (\frac{\sigma^2}{\tau^2}\mathbf{I} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$
$$\Sigma_{\boldsymbol{\beta}|\mathbf{X},\mathbf{y}} = (\tau^{-2}\mathbf{I} + \mathbf{X}^\top\mathbf{X}/\sigma^2)^{-1}$$

Obviously, MAP estimate is the same as the ridge estimate with $\lambda = \sigma^2/\tau^2$.

Assume $\beta$ has the prior distribution where the $\beta_j$ 's are independent and each having mean-zero Laplace distribution:

$$f(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau}\right)$$

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) &\propto p(\beta) \cdot p(\mathbf{y} \mid \mathbf{X}, \beta) \\ &\propto \exp\left(-\frac{|\beta_1|}{\tau} - \cdots - \frac{|\beta_p|}{\tau}\right) \exp\left(-(\mathbf{y} - \mathbf{X}\beta)^\top \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - X\beta) - \frac{1}{\tau} \|\beta\|_1\right) \end{aligned}$$

The MAP is the same as

$$\underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \frac{2\sigma^2}{\tau} \|\beta\|_1$$

# Remark

Inference after model selection?

- the usual hypothesis test, confidence intervals become invalid if model selection is not accounted for.
- **post-selection inference** is challenging and remains a very active research area.