Regularization and Variable Selection

Wei Li

Syracuse University

Spring 2021

Regularization and Variable Selection

Motivation

Predictive Accuracy

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

Then Ordinary Least Squares has low bias. But for variance:

- 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
- When n is not much larger than p, LS fit can have high variance and may result in over fitting
- For a truly linear underlying model, the bias of the linear regression estimate is exactly 0, and the variance is $p \cdot \sigma^2/n$

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.

Subset 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 procedure produce a sequence of models of different size k (i.e., different number of predictors k).

Model Selection

Model Selection

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:

- Mallow's C_p, AIC, BIC (more in Sec 7.5)
- adjusted-R^2
- prediction error on the test set
- cross validation; generalized cross validation (GCV) (more in Sec 7.10)

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

- 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
- The residual sum of squares as $\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i \right)^2$

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

- Suppose we can estimate σ^2 using a full model, its estimate is given by $\tilde{\sigma}^2$,

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

- Adjusted $R^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$

Forward Stagewise Linear Regression

Let's assume that the features have been standardized to sample mean 0 and sample variance 1.

Forward Stagewise Linear Regression, henceforth called Stagewise, is an iterative technique that begins with $\hat{\boldsymbol{\mu}} = 0$ and builds up the regression function in successive small steps. If $\hat{\boldsymbol{\mu}}$ is the current Stagewise estimate, let $\mathbf{c}(\hat{\boldsymbol{\mu}})$ be the vector of *current correlations*

$$\hat{\mathbf{c}} = \mathbf{c}(\hat{\boldsymbol{\mu}}) = \mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}})$$

so that $\hat{c}_j$ is proportional to the correlation between covariate $\mathbf{x}_j$ and the current residual vector. The next step of the Stagewise algorithm is taken in the direction of the greatest current correlation,

$$\hat{j} = \text{argmax}\,|\hat{c}_j| \qquad \text{and} \qquad \hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + \epsilon \cdot \text{sign}\left(\hat{c}_{\hat{j}}\right) \cdot \mathbf{x}_{\hat{j}}$$

with ϵ some small constant. "Small" is important here: the "big" choice $\epsilon = |\hat{c}_j|$ leads to a version of Forward Selection algorithm.

Lasso

Lasso

A generalized regularization framework:

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

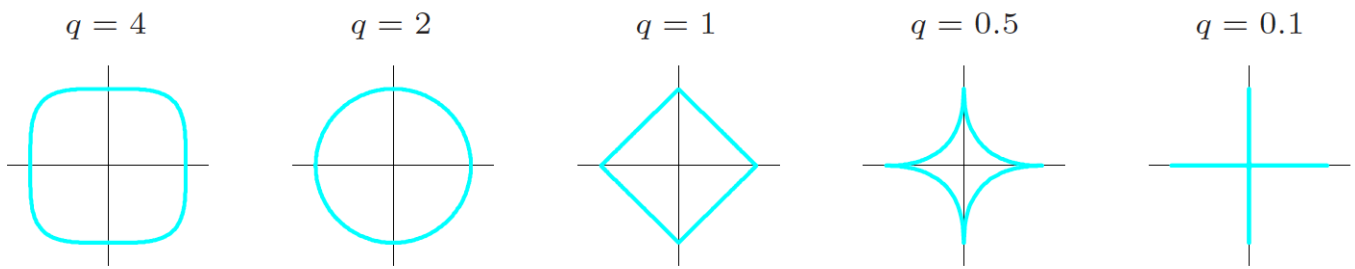
$L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$ is the loss function

- For OLS, $L = \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right]^2$
- For MLE methods, $L = -\log$ likelihood
- Others: hinge loss function (SVM), or exponential loss (AdaBoost)

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

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

- $J_0(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} 1\left(\beta_j \neq 0\right)$. (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, this gives rise to

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

Another way to express the problem is

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

$t \ge 0$.

Lasso

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

In practice, we center predictors and response, and standardize predictors.

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

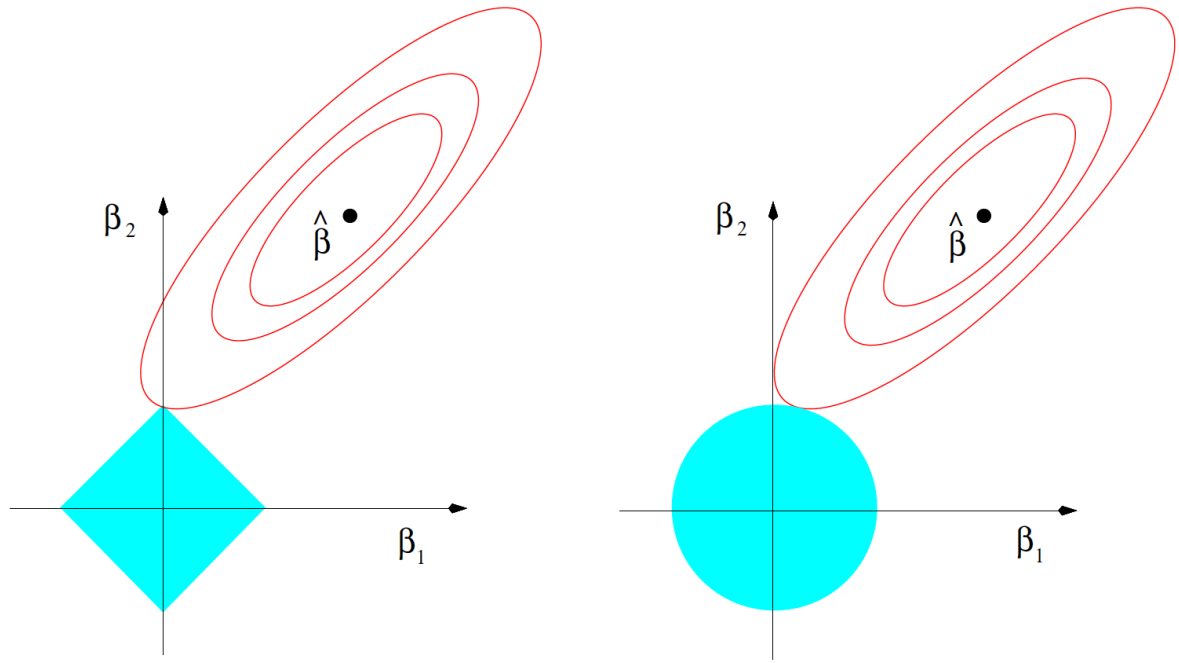
The equivalent optimization problem is

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

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

- Explicitly apply the magnitude constraint on the parameters
- Making t sufficiently small will cause some coefficients to be exactly zero.
- The lasso is a kind of continuous subset selection

Note, t can be chosen between 0 and $\sum_j |\hat{\beta}_j|$ where $\hat{\beta}_j$ are OLS estimates. One can define shrinkage factor $s = t/\sum_j |\hat{\beta}_j|$.



ESL: Fig 3.11 Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

Choose λ

The tuning parameter t or λ should be adaptively chosen to minimize the MSE or PE (prediction error).

- If t is chosen larger than $t_0 = \sum_{j=1}^{p} \left|\hat{\beta}_j^{ols}\right|$, then $\hat{\beta}_j^{\text{lasso}} = \hat{\beta}_j^{ols}$.
- If $t = t_0/2$, the least squares coefficients are shrunk by 50%.
- In practice, we sometimes standardize t by $s = t/t_0$ in $[0, 1]$, and choose the best value s from [0,1] .

The simplest and most popular way to estimate the prediction error.

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

Define the Index function (allocating memberships)

$$\kappa : \{1, \ldots, n\} \to \{1, \ldots, K\}$$

The K -fold cross-validation estimate of prediction error (PE) is

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

Typical choice: $K = 5, 10$

For Lasso, choose a sequence of λ values, say $\{\lambda_1, \ldots, \lambda_m\}$:

- For each λ, fit the LASSO and denote the solution by $\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}$.
- Compute the $CV(\lambda)$ curve as

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

  which provides an estimate of the test error curve.
- Find the best parameter λ^* which minimizes $CV(\lambda)$.

$$\lambda^* = \underset{\lambda \in \{\lambda_1, \ldots \lambda_m\}}{\text{argmin}} \; \text{CV}\,(\lambda)$$

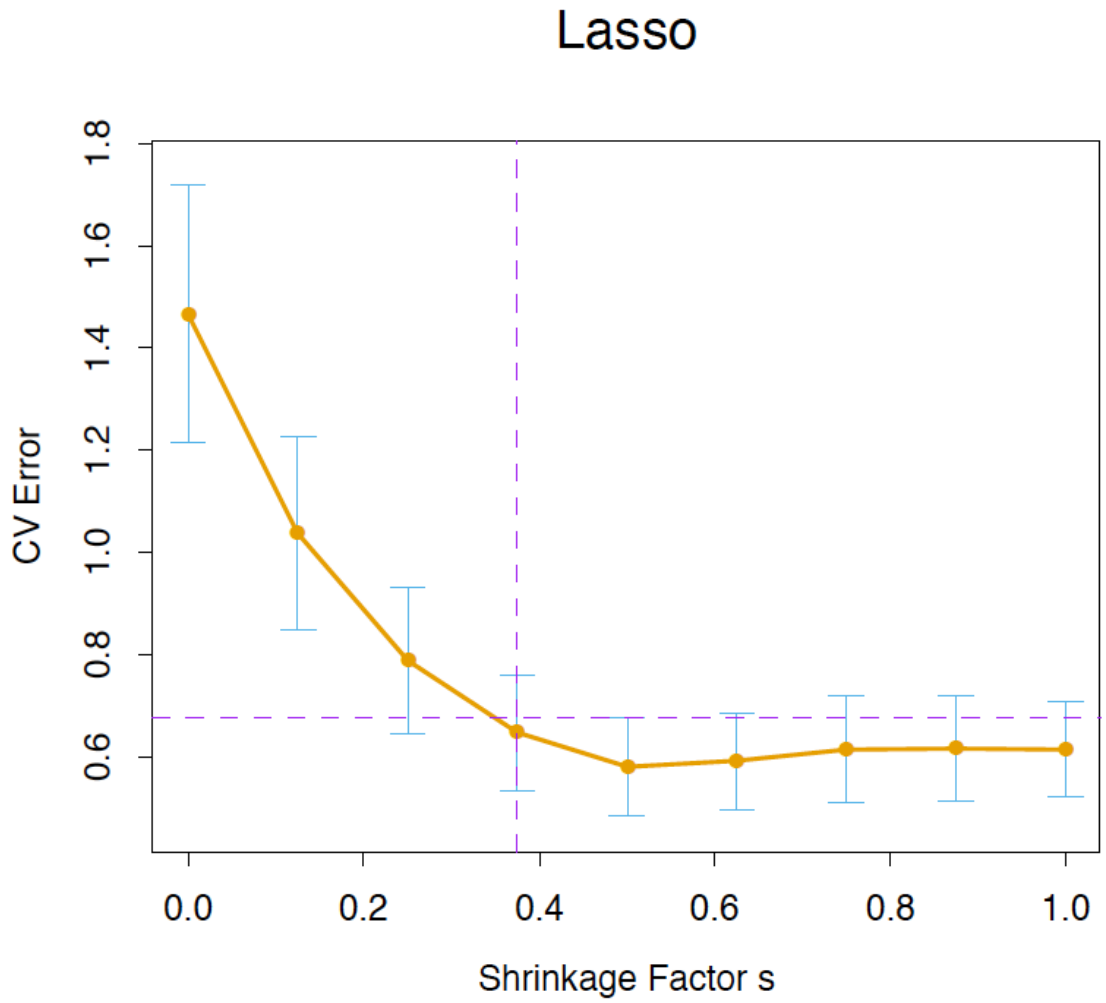
- Fit the final LASSO model with λ^*. The final solution is denoted as $\widehat{\boldsymbol{\beta}}_{\lambda^*}^{\text{lasso}}$

One-standard error rule

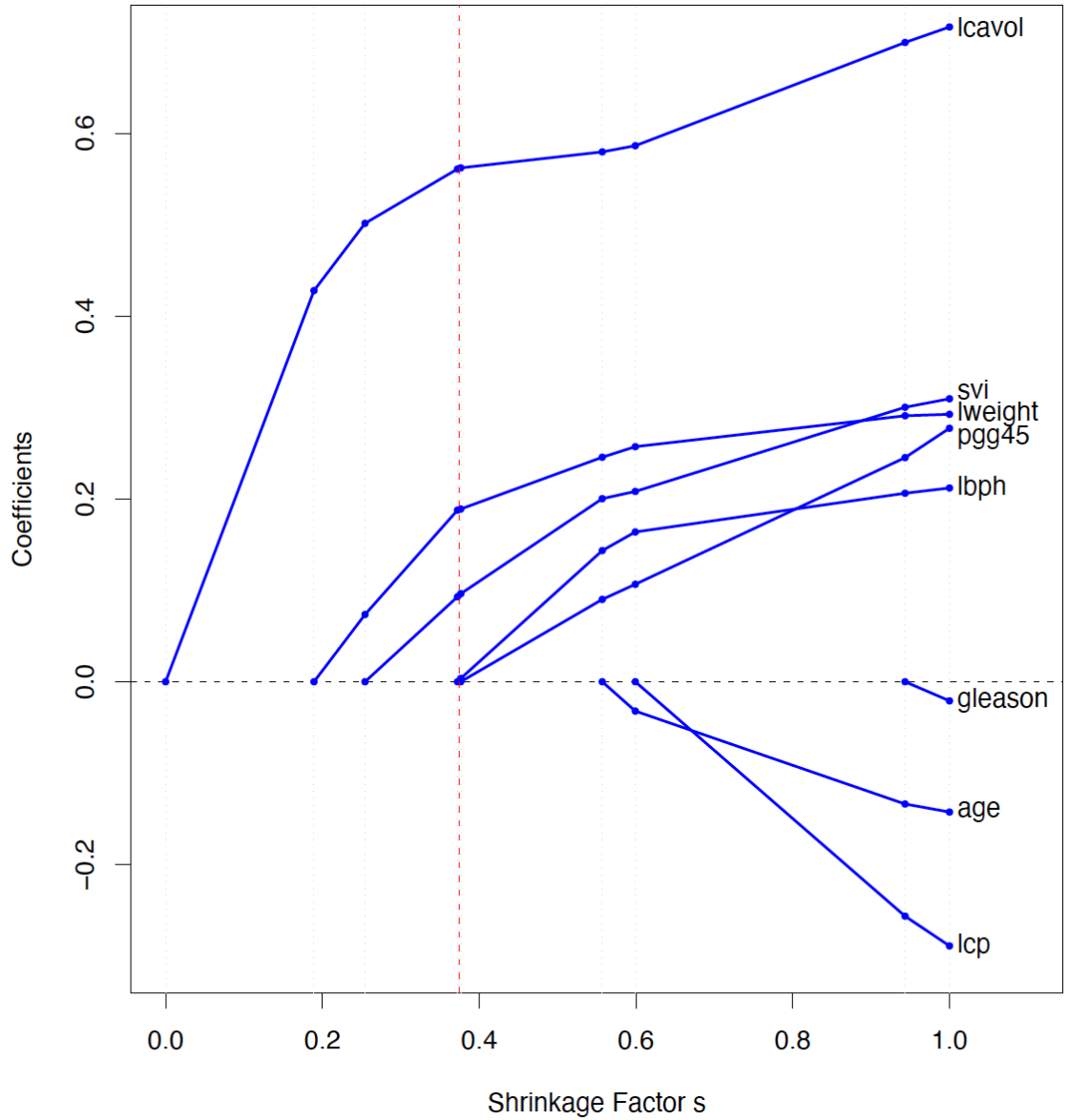
- choose the most parsimonious model with error no more than one standard error above the best error.
- used often in model selection (for a smaller model size).

$$\hat{\lambda} = \max \{\lambda \in \{\lambda_1, \ldots \lambda_m\} : \mathrm{CV}(\lambda) \leq \mathrm{CV}(\lambda^*) + \mathrm{SE}\left(\mathrm{CV}(\lambda^*)\right)\}$$

Prostate Cancer Data Example



ESL Figure 3.7



ESL Figure 3.10

Ridge Regression

Ridge Regression

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

An equivalent way to write the ridge problem is

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

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

The ridge regression solutions is

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

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

Also, there always exists a parameter $\lambda^* > 0$ such that mean-squared-error of is strictly smaller than mean-squared-error of the OLS estimate.

Note the similarity to the ordinary least squares solution, but with the addition of a "ridge" down the diagonal.

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

In the special case of an 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).

Interpreting shrinkage

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

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

- 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.

One can show that

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

$$\mathbf{X}\hat{\beta}^{\text{ols}} = \sum_{j=1}^{p} \mathbf{u}_j \mathbf{u}_j^T \mathbf{y}$$

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

- uses an orthogonal transformation to convert 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 directions**.

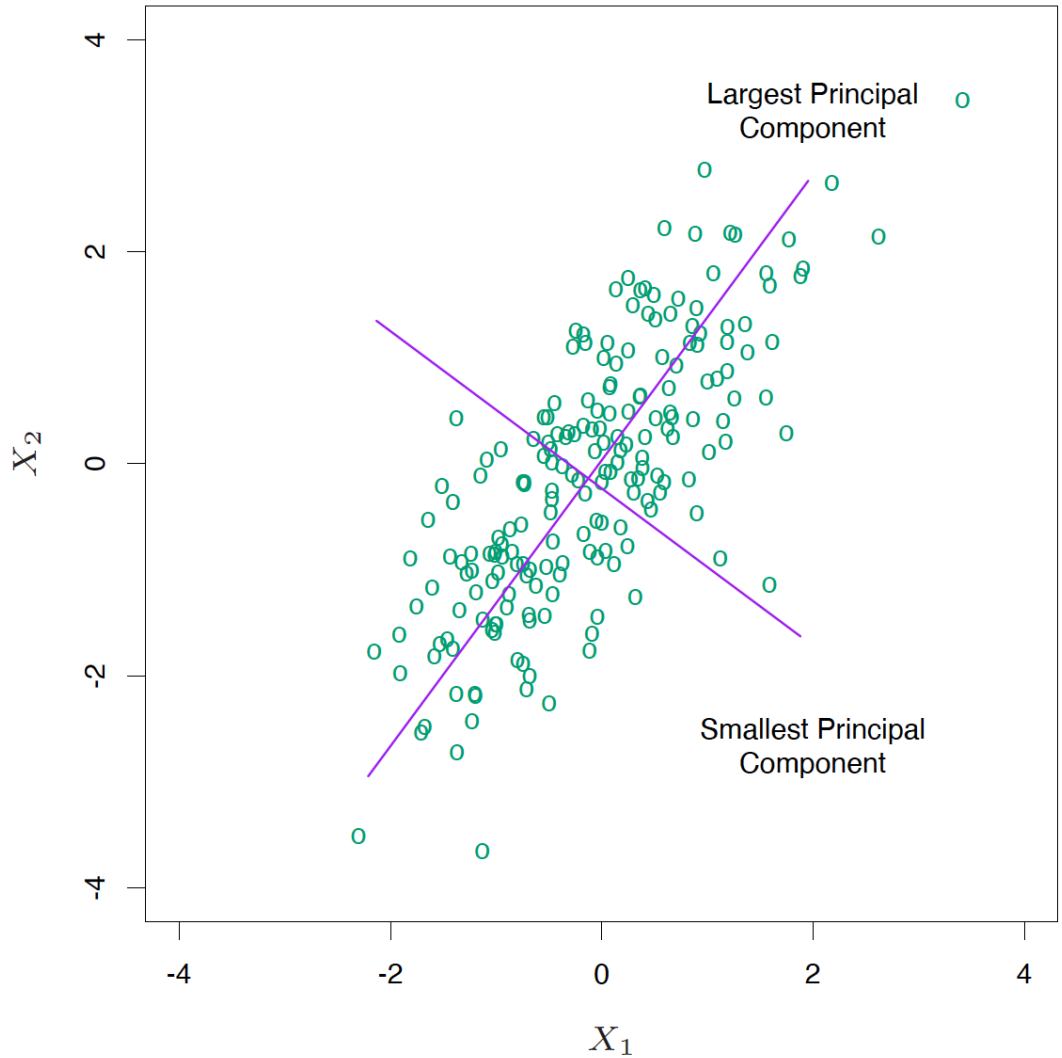
The **first principal component direction** v_1 has the property that $\mathbf{z}_1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$. This sample variance can be seen as

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

and in fact $\mathbf{z}_1 = \mathbf{X}v_1 = \mathbf{u}_1 d_1$. The derived variable $\mathbf{z}_1$ is called the **first principal component** of $\mathbf{X}$.

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

Hence the small singular values d_j correspond to directions in the column space of $\mathbf{X}$ having small variance, and ridge regression shrinks these directions the most.



ESL Figure 3.9

Degrees of Freedom

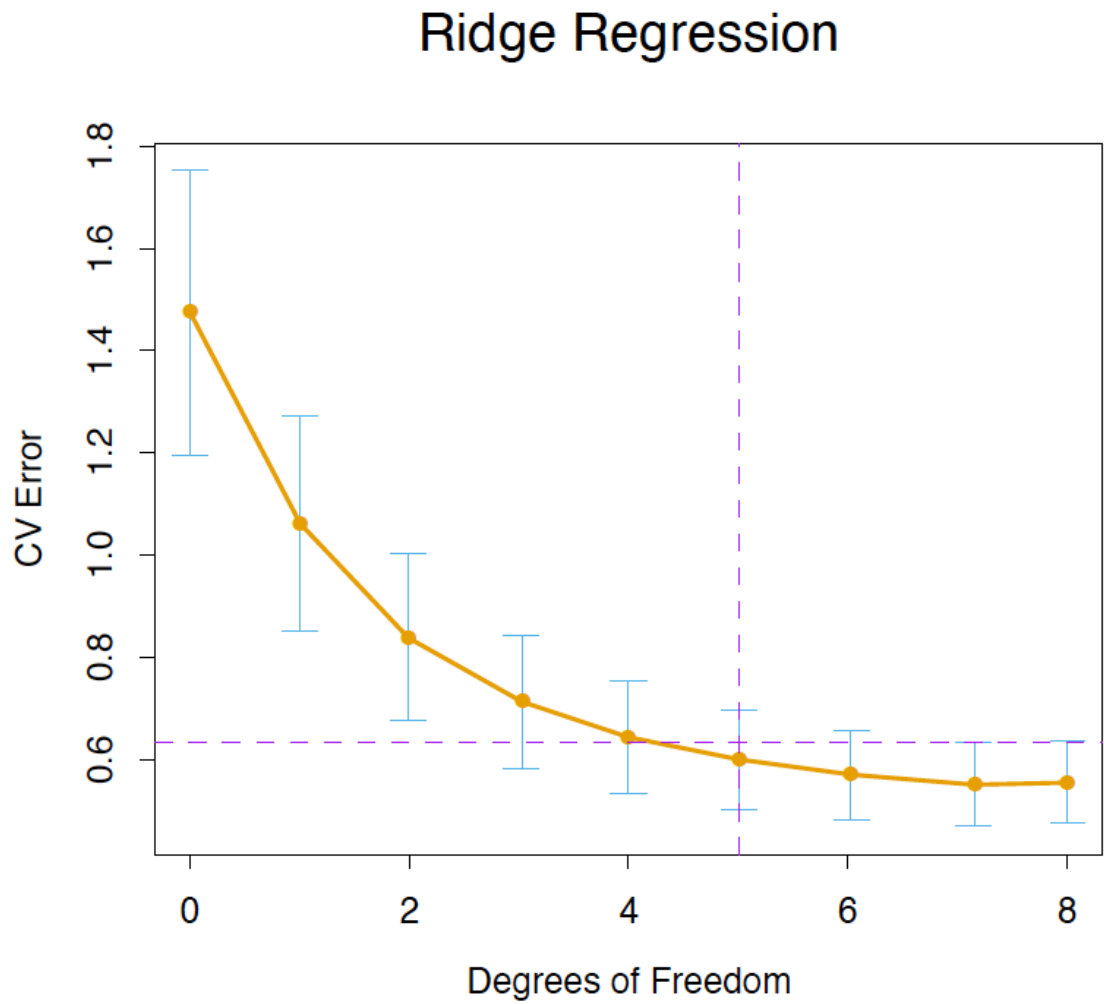
For the ridge estimator, an unbiased estimate for its df (details in Sec 7.6) is

$$\widehat{df}\left(\widehat{\boldsymbol{\beta}}^{\text{ridge}}\right) = \text{trace}\left\{\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I}_d\right)^{-1}\mathbf{X}^T\right\}$$
$$= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

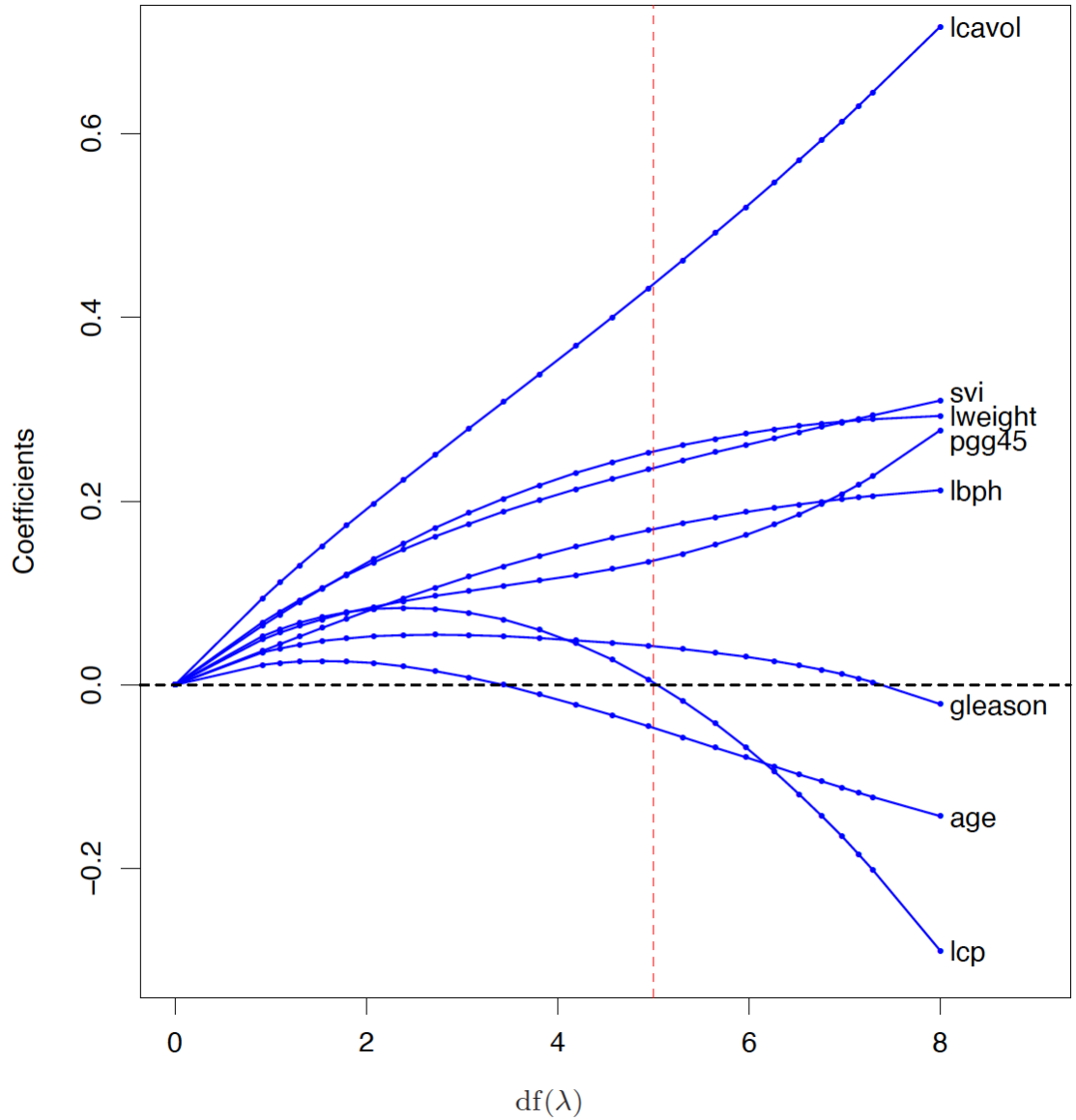
Note:

- $df(\lambda)$ is monotone decreasing functions of λ, called effective d.f.
- $df(\lambda) = p$ when $\lambda = 0$; $df(\lambda) \to 0$ if $\lambda = 0$

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 $\mathbf{X}^T\mathbf{X}$ is not invertible. For any design matrix $\mathbf{X}$, the quantity $\mathbf{X}^T\mathbf{X} + \lambda I_p$ is always invertible; thus, there is always a unique solution $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$
- There always exists a λ such that the MSE of $\boldsymbol{\beta}^{\text{ridge}}$ 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.

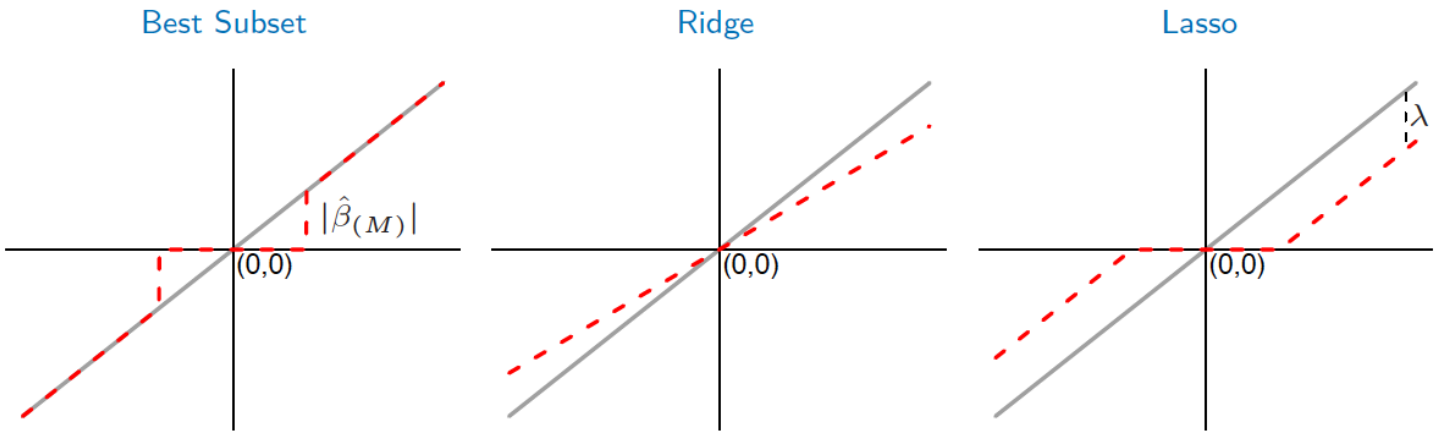
A comparison for orthonormal design: $\mathbf{X}^T\mathbf{X} = I$

- The solution to the Lasso problem is

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

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

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.

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 Σ_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.

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

Here $\mathcal{S}$ indexes the subset of features with non-zero coefficients in the true underlying model, and $\mathbf{X}_{\mathcal{S}}$ are the columns of $\mathbf{X}$ corresponding to those features. Similarly defined for $\mathcal{S}^c$ and $\mathbf{X}_{\mathcal{S}^c}$

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"
- Coordinate Descent Algorithm (CDA; Friedman et al. 2010)
  - designed for GLM
  - suitable for ultra-high dimensional problems
  - R package "glmnet"

Coordinate descent for lasso

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

More general, for $f(x) = g(x) + \sum_{i=1}^n h_i(x_i)$ (with g convex, differentiable and each h_i convex) we can use coordinate descent to find a minimizer: start with some initial guess $x^{(0)}$, and repeat

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

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
    - $z_j = 2 \sum_{i=1}^n (x_{ij})^2$
    - $\rho_j = 2 \sum_{i=1}^n x_{ij}[y_i - \beta^T x_i + \beta_j x_{ij}]$
    - $\beta_j = \text{soft}(\frac{\rho_j}{z_j}; \frac{\lambda}{z_j})$
- until convergence

Beyond Lasso and Remarks

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_ℓ the number in group ℓ.

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

where

- $\lambda \geq 0$ is a tuning parameter.
- $\left\| \boldsymbol{\beta}_j \right\| = \sqrt{\boldsymbol{\beta}_j^T \boldsymbol{\beta}_j}$ 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.

Elastic net (Zhou and Hastie 2005)

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

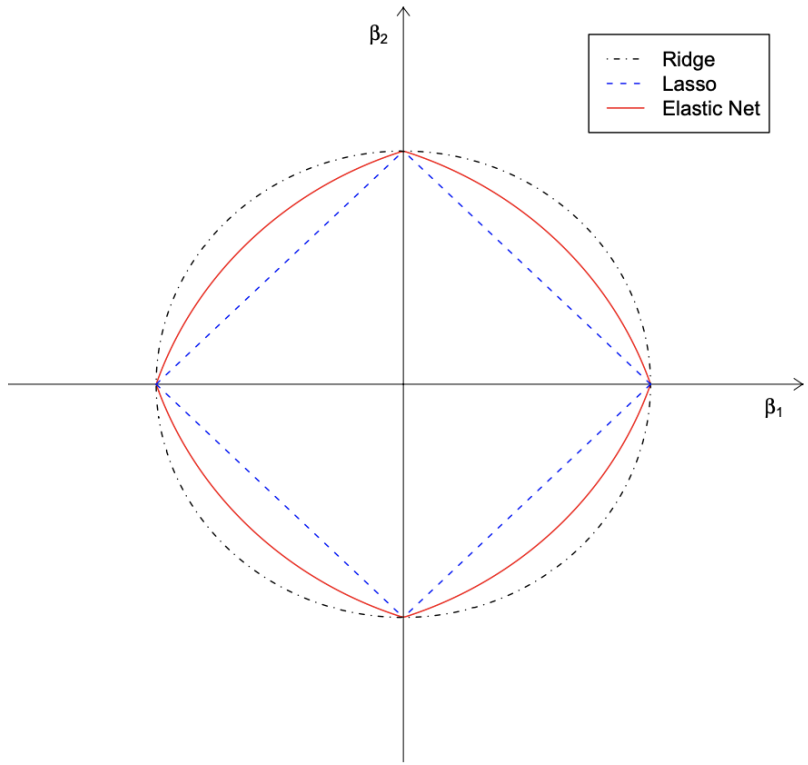


Figure: Elastic net penalty

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

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:

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

.

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.

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

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

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

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

Assume β has the prior distribution where the β_j 's are independent and each having mean-zero Laplace distribution:

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

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

The MAP is the same as

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

Remark

Inference after model selection?

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