Regularization Technique for Linear Regression: Ridge Regression and LASSO

Xiang Zhou

School of Data Science
Department of Mathematics
City University of Hong Kong

Preface

Focus point of this part is still the bias-variance tradeoff by introducing the shrinkage methodology to allow the **biased** estimator. The linear regression problem here is used as a second example [1]of **model assessment and model selection**. (Chapter 7 [ESL])

Not intended to cover at all: the optimization theory and numerical methods for the ridge regression and LASSO, which is very important in practice but more close to the optimization field.

[1]the first is KNN

Revisit the Bias-Variance tradeoff

For a model fit $\hat{f}_{\mathtt{D}}(x)$ based on the data $\mathtt{D} = (\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is the design matrix and $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ is the response data, a good measure of the quality of this model at a new test input $x_0 \in \mathbb{R}^p$ is the mean square error (MSE). Let $f(x) = x^{\mathsf{T}}\beta$ be the true value of the output at the point x, then

$$\begin{aligned}\mathsf{MSE}(\hat{f}_{\mathtt{D}}(x_0)) &:= \mathbb{E}_{\mathtt{D}}(\hat{f}_{\mathtt{D}}(x_0) - f(x_0))^2 \\ &= \mathsf{Var}_{\mathtt{D}}(\hat{f}_{\mathtt{D}}(x_0)) + \Big(\mathbb{E}_{\mathtt{D}}\, \hat{f}_{\mathtt{D}}(x_0) - f(x_0)\Big)^2 .\end{aligned}$$

- Typically, when bias is low, variance will be high and vice-versa. Choosing estimators often involves a tradeoff between bias and variance.
- So far, OLS estimator $\hat{\beta}^{OLS} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ is unbiased and also has the minimal MSE if restricted to be *unbiased* and linear in $\mathbf{y}$.

For the OLS fit, its MSE (written with dependency on x_0) is

$$
\begin{aligned}
\mathsf{MSE}^{OLS}(x_0) &= \mathsf{Var}_{\mathbf{y}}(x_0^\mathsf{T}\hat{\beta}^{OLS}) \\
&= \mathsf{Var}_{\mathbf{y}}(x_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}) \\
&= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}x_0\,\mathbb{V}(\mathbf{y})x_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X} \\
&= \sigma_\varepsilon^2\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}x_0x_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \\
&= \sigma_\varepsilon^2\left\|\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}x_0\right\|_2^2
\end{aligned}
$$

Here $\mathbf{X}$ is treated as deterministic or say that we work on the condition variance of given $\mathbf{X}$.

Exercise

In the above, $x_0 \in \mathcal{X}$ is arbitrary. Now let x_0 take to be each of training sample value x_i, $1 \le i \le N$, show that the in-sample error, which is defined by $\frac{1}{N}\sum_{i=1}^N$ MSE$^{OLS}(x_i)$, equals to $\frac{p}{N}\sigma_\varepsilon^2$ where p is the dimension of $\mathcal{X}$. (Equation (7.29) in [ESL])

There can be biased estimators with smaller MSE. The following property is quite general.

Exercise

Define a biased estimate of the coefficient β in the following special form

$$\widetilde{\beta} = (1+\alpha)\hat{\beta}$$

with a scalar α where $\hat{\beta}$ is an unbiased estimate. Calculate the MSE of $\widetilde{\beta}$ and find a condition that MSE$(\widetilde{\beta}) <$ MSE$(\hat{\beta})$.

- Generally, by regularizing the estimator in some way, its variance will be reduced; if the corresponding increase in bias is small, this will be worthwhile.
- Examples of regularization: subset selection (forward, backward, all subsets)[1]; ridge regression, lasso

[1] read [ISL][ESL] by yourself; not to cover in lecturing

Ridge Regression

Regression in high dimension

- High dimensional problem: the input dimension $d = p + 1$ is close or greater than the number of samples n. [1]
- When the design matrix $\mathbf{X}$ is high-dimensional, the covariates (the columns of $\mathbf{X}$) are super-collinear. *collinearity* in regression analysis refers to the event of two (or multiple) covariates being highly linearly related.
- In OLS estimate, the dimension of the subspace X may be less than d and then the matrix inverse $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ does not exist or $\mathbf{X}^\mathsf{T}\mathbf{X}$ is close to singular even if invertible[2].

---

[1] we switch symbols both d and $p+1$ from now. The notations for design matrix and response variable are changed to bold font.

[2] Numerical linear algebra uses the **condition number** (defined as the ratio of largest to the smallest eigenvalue) to represent the illness of the problem: the larger, the worse.

Solution of ridge regression

Ridge Regression solves

$$\min_{\beta} \quad \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

The solution is

$$\hat{\beta}^{\lambda} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Inclusion of λ makes problem non-singular even if $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is not invertible: This was the original motivation for ridge regression (Hoerl and Kennard, 1970)

Note $\lambda = 0$ gives the ordinary least squares estimator, and if $\lambda \to \infty$ then $\hat{\beta}_{\lambda} \to 0$. In general, with a good choice of λ, $\hat{\beta}_{\lambda}$ is a biased estimator that may have smaller mean squared error than the least squares estimator

Eigenvalue shrinkage in ridge regression

Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$ is a singular value decomposition of $\mathbf{X}$: $\mathbf{D}$ is the $(p+1)\times(p+1)$ diagonal matrix consisting of singular values $d_0 \geq d_1 \geq \ldots \geq d_p$. $\mathbf{U}$ and $\mathbf{V}$ are $n \times (p+1)$ and $(p+1)\times(p+1)$ matrices, respectively [1]. Then $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\mathsf{T}$ and the ridge regression solution

$$
\begin{aligned}
\hat{\beta}^\lambda &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^\mathsf{T} + \lambda\mathbf{V}\mathbf{V}^\mathsf{T})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}\,\mathsf{diag}\left\{\frac{d_j}{d_j^2+\lambda}\right\}\mathbf{U}^\mathsf{T}\mathbf{y},
\end{aligned}
$$

which is well defined for any d_j when λ is strictly positive.

[1] The column space of $\mathbf{U}$ is the column space of $\mathbf{X}$ in $\mathbb{R}^n$ and the column space of $\mathbf{V}$ is the row space of $\mathbf{X}$ in $\mathbb{R}^{p+1}$. $\mathbf{U}^\mathsf{T}U = \mathbf{I}_n$ and $\mathbf{V}\mathbf{V}^\mathsf{T} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_{p+1}$.The number of nonzero $\{d_j\}$ is the rank of $\mathbf{X}$.

Let $\widetilde{\beta} = \mathbf{V}^\mathsf{T}\hat{\beta}$ and $\widetilde{\mathbf{y}} = \mathbf{U}^\mathsf{T}\mathbf{y}$. Then

$$\widetilde{\beta}^\lambda = \mathbf{V}^\mathsf{T}\hat{\beta}^\lambda = \mathrm{diag}\left\{\frac{d_j}{d_j^2+\lambda}\right\}\widetilde{\mathbf{y}}$$

If $\mathbf{D}$ is nonsingular, then $\widetilde{\beta}^{OLS} = \widetilde{\beta}^0 = \mathbf{D}^{-1}\widetilde{\mathbf{y}}$ exists. So

$$\widetilde{\beta}^\lambda = \mathbf{D}_\lambda\mathbf{D}^{-1}\widetilde{\mathbf{y}} = \mathbf{D}_\lambda\widetilde{\beta}^{OLS} \quad \text{where} \quad \mathbf{D}_\lambda := \mathrm{diag}\left\{\frac{d_j^2}{d_j^2+\lambda}\right\} \le \mathbf{I}$$

$$\boxed{\widetilde{\beta}_j^{ridge} = \frac{d_j^2}{d_j^2+\lambda}\widetilde{\beta}_j^{OLS}}$$

Exercise

Find the bias, the variance and the MSE for the transformed ridge coefficient $\widetilde{\beta}^\lambda$ in terms of $\beta, \mathbf{X}, \mathbf{y}, \sigma^2$. Find the optimal λ in theory that minimize the MSE.

Smoother matrix and effective degree of freedom

- A smoother matrix S is a linear operator satisfying

$$\mathbf{y} = \mathbf{S}\mathbf{y}$$

  where $\mathbf{S}$ may depend on $\mathbf{X}$, but not $\mathbf{y}$.

- define the **effective degrees of freedom** (or effective number of parameters) for a smoother $\mathbf{S}$:

$$\text{df}(\mathbf{S}) = \mathrm{Trace}(\mathbf{S})$$

Exercise

Ex. 7.4 and 7.5 in [ESL]

Effective degree of freedom

The ridge solution can be rewritten as

$$
\begin{aligned}
\hat{\mathbf{y}}^{ridge} &= \mathbf{X}\hat{\beta}^{ridge} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}\mathbf{V}\mathbf{D}_{\lambda}\mathbf{D}^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{y} \\
&= \mathbf{U}\mathbf{D}_{\lambda}\mathbf{U}^{\mathsf{T}}\mathbf{y} \\
&= \sum_{j=0}^{p}\Big(\frac{d_j^2}{d_j^2+\lambda}\Big)\mathbf{U}_j\mathbf{U}_j^{\mathsf{T}}\mathbf{y},
\end{aligned}
$$

$\mathbf{U}_j$ is the column vector of $\mathbf{U}$. The effective degree of freedom of the ridge regression is

$$
\begin{aligned}
df(\lambda) &:= \mathrm{Trace}\left(\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\right) = \mathrm{Trace}(\mathbf{V}\mathbf{D}_{\lambda}\mathbf{D}^{-1}\mathbf{U}^{\mathsf{T}}) \\
&= \mathrm{Trace}(\mathbf{D}_{\lambda}\mathbf{D}^{-1}) = \sum_{j=0}^{p}\frac{d_j^2}{d_j^2+\lambda}
\end{aligned}
$$

decreases from $p+1$ to 0 as λ increases from 0 to ∞.

Ridge regression equivalent form

$$\hat{\beta}^{\lambda} = \underset{\beta}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \,,$$

$$\hat{\beta}^{t} = \underset{\|\beta\|_2^2 \le t}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

One can prove that there exists a bijection between λ and t.

Exercise

Find the bijection between two positive scalars λ and t if two vectors β^{λ} and β^{t} are the same.

The least square error as the $\mathbf{D}$-weighted distance to $\hat{\beta}^{OLS}$ in the transformed coordinate system

$$\begin{aligned}
&\|\mathbf{y}-\mathbf{X}\beta\|_2^2 \\
&= \left\|\mathbf{U}\mathbf{U}^\mathsf{T}\mathbf{y}-\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}\beta\right\|_2^2 = \left\|\mathbf{U}^\mathsf{T}\mathbf{y}-\mathbf{D}\mathbf{V}^\mathsf{T}\beta\right\|_2^2 \\
&= \|\widetilde{\mathbf{y}}-\mathbf{D}\widetilde{\beta}\|_2^2 = \|\mathbf{D}(\widetilde{\beta}^{OLS}-\widetilde{\beta})\|_2^2
\end{aligned}$$

For the constraint $\|\beta\|_2 = \|\mathbf{V}^\mathsf{T}\beta\| = \|\widetilde{\beta}\| \leq t$, one can show the inequality problem $\min_{\|\widetilde{\beta}\|_2^2 \leq t} \|\widetilde{\mathbf{y}}-\mathbf{D}\widetilde{\beta}\|_2^2$ actually attains the equality constraint for this case. The Lagrangian function is $\|\widetilde{\mathbf{y}}-\mathbf{D}\widetilde{\beta}\|_2^2 + \mu(\|\widetilde{\beta}\|_2^2 - t)$ So, KKT gives $\mathbf{D}^2\widetilde{\beta} - \mathbf{D}\widetilde{\mathbf{y}} + \mu\widetilde{\beta} = 0$, i.e., $\widetilde{\beta} = (\mathbf{D}^2+\mu\mathbf{I})^{-1}\mathbf{D}\widetilde{\mathbf{y}}$. So, μ is the same as λ. The equality constratin $\|\widetilde{\beta}\| = t$ determines uniquely $\beta = \mu = \mu(t)$.

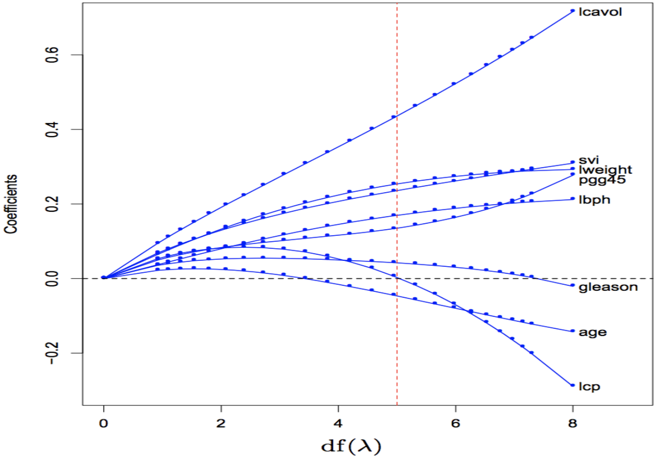


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.* [1]

[1] solution path of ridge regression, from [ESL].

LASSO

Bridge estimators

(Frank and Friedman, 1993)
With $L_r(\beta) = \sum_{j=0}^{p} |\beta_j|^r$,

$$\hat{\beta}^{bridge} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda L_r(\beta)$$

- $L_0(\beta) = \sum_{j=0}^{p} I(\beta_j \neq 0)$; (Hard thresholding)
- $L_1(\beta) = \sum_{j=0}^{p} |\beta_j|$; (Lasso)
- $L_2(\beta) = \sum_{j=0}^{p} \beta_j^2$; (Ridge regression)
- $L_\infty(\beta) = \max_j \ |\beta_j|$.

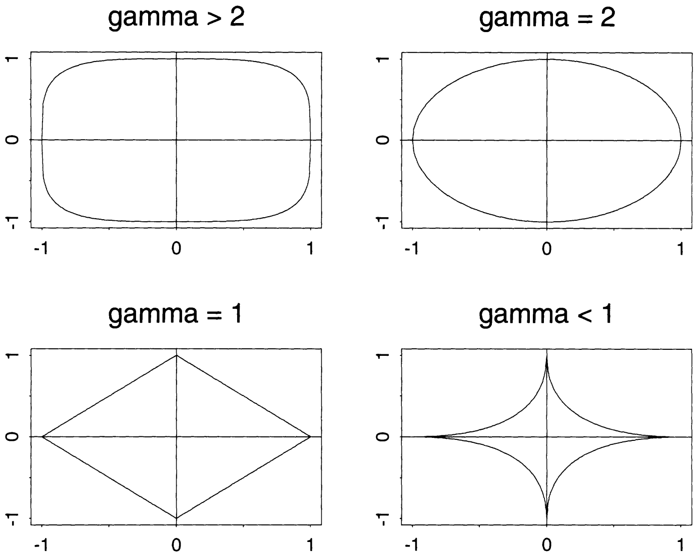


Figure 1. Constrained Areas of Bridge Regressions with t = 1.

Least Absolute Shrinkage and Selection Operator (Lasso)

Tibshirani (Journal of the Royal Statistical Society 1996) introduced the LASSO.

Lasso estimator: let $r = 1$ in bridge estimator

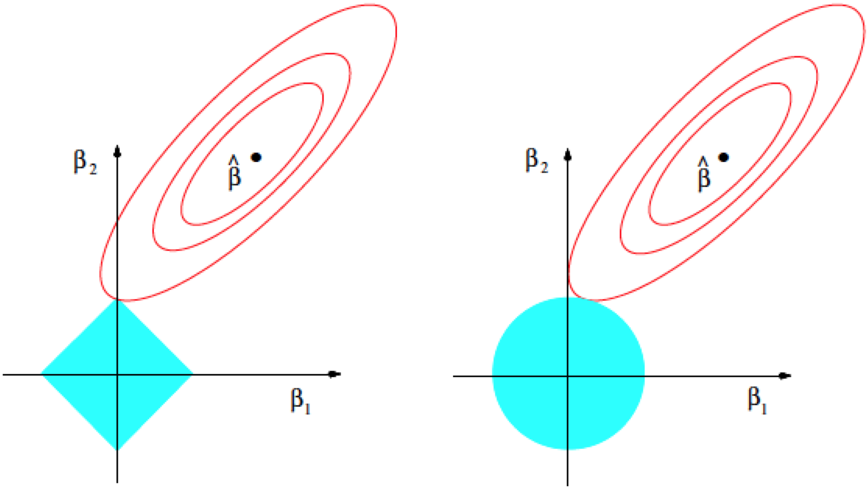
$$\hat{\beta}^{\lambda} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

$$\hat{\beta}^{s} = \underset{\|\beta\|_1 \leq s}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2,$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.
Again there exists a bijection between λ and s.

Sparse solution

- Due to the nature of the l_1-norm constraint, if t is small enough some coefficients of the lasso solution become *exactly* zero.
- The elliptical contour is likely to hit the corner of the polytope, corresponding to sparse $\hat{\beta}$. Variable selection: drop the features with $\hat{\beta}_j = 0$.
- The l_r regularization results in sparsity when $0 \leq r \leq 1$, and is convex when $1 \leq r < \infty$.
- **Lasso is sparse and convex.** Original implementation involves quadratic programming techniques from convex optimization
- Efron et al. (Annals of Statistics 2004) proposed **LARS** (**least angle regression**), which computes the LASSO path efficiently
  - Interesting modification called is called forward stagewise
  - In many cases it is the same as the LASSO solution
  - Forward stagewise is easy to implement: http://www-stat.stanford.edu/~hastie/TALKS/nips2005.pdf



$\hat{\beta}$ in the center is the $\hat{\beta}^{OLS}$

Ridge vs Lasso

Consider a simple but illuminating example: Show the solutions

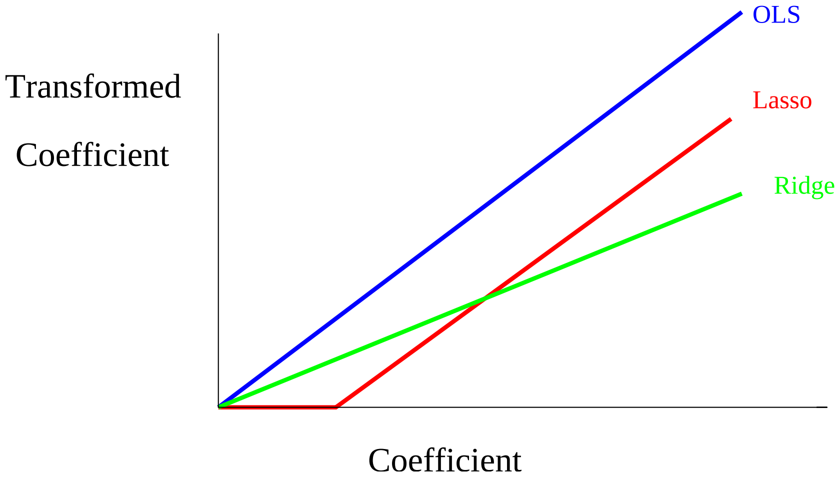
$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\beta - \hat{\beta}\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\beta - \hat{\beta}\|_2^2 + \lambda \|\beta\|_1 \right\}$$

are

- $\hat{\beta}_j^{ridge} = \hat{\beta}_j/(1+\lambda)$
- $\hat{\beta}_j^{lasso} = \mathsf{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$

visualization of $\hat{\beta}^{ridge}$ and $\hat{\beta}^{lasso}$ from desmos.com webpage
Warning: taking the traditional derivative for lasso objective function is the wrong way!
Ridge regression shrinks the $\hat{\beta}$ in all components/directions. Lasso translates them towards zero by a constant, truncating at zero.


OLS
Lasso
Ridge
Transformed
Coefficient
Coefficient

Solution path

- Clearly, the lasso solution $\hat{\beta}^{lasso}$ changes with the value of λ
- Even better, $\hat{\beta}^{lasso}(\lambda)$ is a piecewise linear function of λ
- The path algorithm starts at $\lambda = \infty$ or $s = 0$, and traces the solution path by continuously changing λ [1]. Each new λ-solution is computed successively by solving the KKT conditions with a good initial guess set as the precise solution for a neighboring old λ.
- The key is to find the turning knots $\lambda_1, \ldots, \lambda_T$
- An interesting reading: Efron et al. (AOS; 2003)
- Read the reference Regularization Paths for Generalized Linear Models via Coordinate Descent
- Path algorithms are available for many methods, such as fused lasso, trend filtering, locally adaptive regression splines, SVMs, 1-norm SVMs, relaxed maximum entropy method ...

[1] homotopy method

The motivation for the Lasso came from an interesting proposal of Breiman (1993). Breiman's non-negative Garotte minimizes

$$\min_{c_j \geq 0, \forall j} \ \frac{1}{2} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} c_j$$

and then $\hat{\beta}_j^{ng} = \hat{c}_j \hat{\beta}_j$.

Exercise

Show when $\mathbf{X}^T \mathbf{X} = n\mathbf{I}$, the non-negative Garotte estimator is

$$\hat{\beta}_j^{ng} = \left(1 - \frac{\lambda}{2\hat{\beta}_j^2}\right)_+ \hat{\beta}_j.$$

- It shrinks small $|\hat{\beta}_j|$ to zero. It is almost unbiased for large $|\hat{\beta}_j|$.
- Garotte starts with the OLS estimates and shrinks them by non-negative factors whose sum is constrained.
- In contrast, Lasso avoids the explicit use of the OLS estimates.
- Lasso is also closely related to the wavelet soft-thresholding method by Donoho and Johnstone (1994), and boosting method.

One can just focus on the component-wise formulation

$$\min_{\beta} \; \frac{1}{2}(z_j - \beta_j)^2 + J(|\beta_j|)$$

- Hard thresholding: $J(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda)$, then[1]

$$\hat{\beta}_j = z_j I(|z_j| > \lambda).$$

- Soft thresholding (Lasso): $J(|\beta_j|) = \lambda|\beta_j|$, then

$$\hat{\beta}_j = \mathsf{sign}(z_j)(|z_j| - \lambda)_+.$$

[1] the proof is exercise

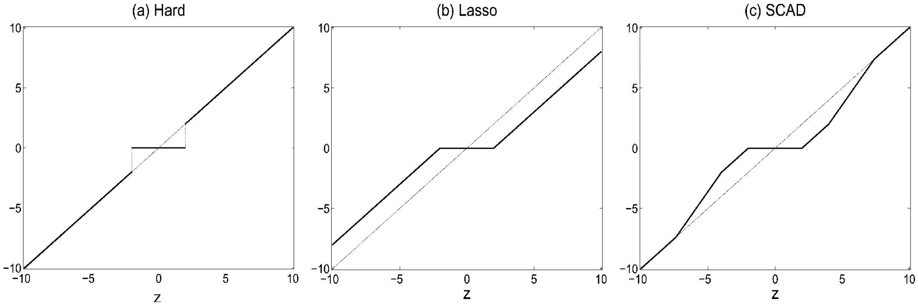
SCAD (Fan and Li, 2001; JASA)

$$\min_{\beta} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} q_\lambda(|\beta_j|),$$

where

$$q_\lambda(\beta) = \begin{cases} \lambda\beta & \text{if } |\beta| \leq \lambda, \\ -\frac{(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda, \\ \frac{(a+1)\lambda}{2} & \text{if } |\beta| > a\lambda. \end{cases}$$

- Similar to the truncated lasso penalty, but it is smooth everywhere.



Reference: Antoniadis and Fan (2001); Fan and Li (2001)

Revisit Bias-Variance (again !)

- In the beginning, We calculated the MSE for ordinary least square where the bias vanishes.
- We now focus on the ridge regression, $\hat{\beta}^{\lambda} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$, where $\lambda > 0$ is the penalty parameter. The corresponding MSE for $\hat{\beta}^{\lambda}$ at the new testing point x_0 then is written as [1]

$$
\begin{aligned}
\mathsf{MSE}^{\lambda}(x_0) &= \mathsf{Var}_{\mathbf{y}}(x_0^{\mathsf{T}}\hat{\beta}^{\lambda}) + \Big(\mathbb{E}_{\mathbf{y}}(x_0^{\mathsf{T}}\hat{\beta}^{\lambda}) - f(x_0)\Big)^2 \\
&= \mathsf{Var}_{\mathbf{y}}(x_0^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}) + \Big(x_0^{\mathsf{T}}\,\mathbb{E}_{\mathbf{y}}(\hat{\beta}^{\lambda}) - f(x_0)\Big)^2 \\
&= \sigma_{\varepsilon}^2\,\|\mathbf{h}_{\lambda}x_0\|_2^2 + \Big(x_0^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta - x_0^{\mathsf{T}}\beta\Big)^2
\end{aligned}
$$

where $\mathbf{h}_{\lambda} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$. Exercise: with the aid of SVD, find the minimum point λ for MSE^{λ}.

[1] again, the training data's $\mathbf{X}$ part is fixed or conditioned.

Decomposition of average squared bias

- Note that we have <u>assumed</u> that the ground truth is a linear model $f(x) = x^\mathsf{T}\beta$.
- From now we do not assume f is linear and it could be a nonlinear function for a general consideration. This is a more realistic setting.
- The additive error model $Y = f(X) + \varepsilon$ is still assumed for data.
- The best-fitting approximation in the *linear model class* [1] is given by

$$\beta^* = \operatorname*{argmin}_{\beta} \mathbb{E}_X (f(X) - X^\mathsf{T}\beta)^2$$

- Note that β_* satisfies the normal equation:

$$\mathbb{E}(X^\mathsf{T}X)\beta_* = \mathbb{E}[f(X)X]$$

[1] In other words, $x \to x^\mathsf{T}\beta^*$ is $f_{\mathcal{H}}(x)$ defined in Topic 1 ($\mathcal{H}$ is hypothesis space)

Decomposition of average (squared) bias

- We still consider the MSE$^\lambda$ for the ridge regression. The variance part $\mathrm{Var}(x_0^\mathsf{T}\hat\beta^\lambda)$ is unchanged. Now the squared bias becomes $\left(x_0^\mathsf{T}\,\mathbb{E}_{\mathbf{y}}(\hat\beta^\lambda) - f(x_0)\right)^2 = \left(x_0^\mathsf{T}\,\mathbb{E}_{\mathbf{y}}(\hat\beta^\lambda) - x_0^\mathsf{T}\beta_* + x_0^\mathsf{T}\beta_* - f(x_0)\right)^2 =$ $\left(x_0^\mathsf{T}\,\mathbb{E}_{\mathbf{y}}(\hat\beta^\lambda) - x_0^\mathsf{T}\beta_*\right)^2 + \left(x_0^\mathsf{T}\beta_* - f(x_0)\right)^2 +$ $2\left(x_0^\mathsf{T}\,\mathbb{E}_{\mathbf{y}}(\hat\beta^\lambda) - x_0^\mathsf{T}\beta_*\right)\left(x_0^\mathsf{T}\beta_* - f(x_0)\right)$
- Taking expectation for $x_0 \sim X$ and using the normal equation, we have the average squared bias is

$$
\begin{aligned}
&\mathbb{E}_{x_0}\left(x_0^\mathsf{T}\,\mathbb{E}_{\mathbf{y}}(\hat\beta^\lambda) - f(x_0)\right)^2 \\
=&\underbrace{\mathbb{E}_{x_0}\left(x_0^\mathsf{T}(\mathbb{E}_{\mathbf{y}}\,\hat\beta^\lambda - \beta_*)\right)^2}_{\text{Ave (Estimation Bias}^2\text{)}} + \underbrace{\mathbb{E}_{x_0}\left(x_0^\mathsf{T}\beta_* - f(x_0)\right)^2}_{\text{Ave (Model Bias}^2\text{)}}
\end{aligned}
$$

This is equation (7.14) in [ELS].

Remarks

- Model Bias, by definition, involves the ground truth, which is not accessible; Estimation Bias, by definition, involves β^*, which is not accessible either.
- However, this decomposition is conceptually inspiriting to have Model Bias.
- The Model Bias corresponds to the approximation error in Topic 1 ($f_{\mathcal{H}}$ vs f^*) and the Estimation Bias is similar to the sampling error in Topic 1 ($\hat{f}_{\mathtt{D}}$ vs $f_{\mathcal{H}}$). In approximation theory viewpoint, the variance $\mathrm{Var}_{\mathtt{D}}\, \hat{f}_{\mathtt{D}}$ is not considered.
- Ridge method and LASSO further restricts the linear model $\mathcal{H}_{linear}$ to a smaller set in the form $\mathcal{H}_\lambda := \left\{ f : f(x) = \beta^\mathsf{T} x, \|\beta\| \leq \lambda \right\}$; they affect the estimation bias (and the variance of the prediction). The improvement of Model Bias needs go from linear to nonlinear if the ground truth is far away from being linear.