High-Dimensional Linear regression and Lasso

$$Y = \beta' X + \varepsilon \quad , \quad \varepsilon \perp X$$

$\dim(X) = p$

* High dimensionality comes from
  1. many regressions eg. growth analysis, health record
  2. transformation to dictionary

* Overfitting: Signal and noise in population model
  OLS, by design, seek the best in sample fitting
  too many regressors will fit the noise,
  resulting in poor out-of-sample prediction

* How to avoid overfitting in high dimension?
  Regularization: Shrinkage estimation
  正则化

  Ridge estimation

$$\frac{1}{n} \| Y - X\beta \|_2^2 + \lambda \|\beta\|_2^2$$

$$\frac{1}{n} (Y - X\beta)'(Y - X\beta) + \lambda \beta'\beta$$

FOC $-\frac{2}{n} \underset{p\times n}{X'} \underset{n\times 1}{(Y - X\beta)} + 2\lambda\beta = 0$

$$\left(\frac{X'X}{n} + \lambda\right)\beta = \frac{XY}{n}$$

Solve $\hat{\beta} = \left(\frac{X'X}{n} + \lambda\right)^{-1} \frac{X'Y}{n}$

Very close to OLS.

Diagonalize the Gram matrix $\frac{X'X}{n} = UDU'$

$$\frac{X'X}{n} + \lambda = U(D + \lambda)U' = U \begin{pmatrix} d_1+\lambda & & & \\ & d_2+\lambda & & \\ & & \ddots & \\ & & & d_p+\lambda \end{pmatrix} U'$$

$$\left(\frac{X'X}{n} + \lambda\right)^{-1} = U \begin{pmatrix} \frac{1}{d_1+\lambda} & & \\ & \ddots & \\ & & \frac{1}{d_p+\lambda} \end{pmatrix}^{-1} U'$$

Prevent λ_j from being to close to 0.

* Asy analysis is very straightforward under fixed p and $\liminf_{n\to\infty} d_p > 0$

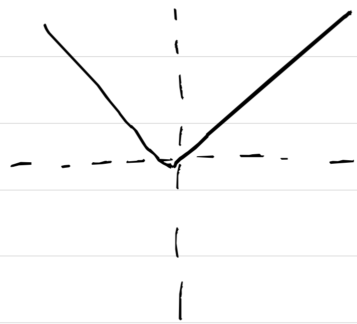
  $\lambda \to 0$ maintains consistency

  Ridge estimator is asy. equivalent to the OLS.

* high dimension anaysis $p/n \to$ const is much more challenging and need advanced math tools.

* Lasso.

$$\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$



45° lines

$$|u| = \begin{cases} u, & \text{if } u > 0 \\ -u, & \text{if } u < 0 \\ 0, & \text{if } u = 0 \end{cases}$$

Subgradient: $\frac{\partial}{\partial u}|u| = \begin{cases} 1, & \text{if } u > 0 \\ -1, & \text{if } u < 0 \\ \text{any number} \in (-1, 1), & \text{if } u = 0 \end{cases}$

heuristic

FOC $\hat{\beta}_j$: $-\frac{1}{n} X_j'(Y - X\hat{\beta}_j) + \lambda \frac{\partial}{\partial \beta_j}|\hat{\beta}_j| = 0$

if $\hat{\beta}_j > 0$, or $\hat{\beta}_j < 0$, then

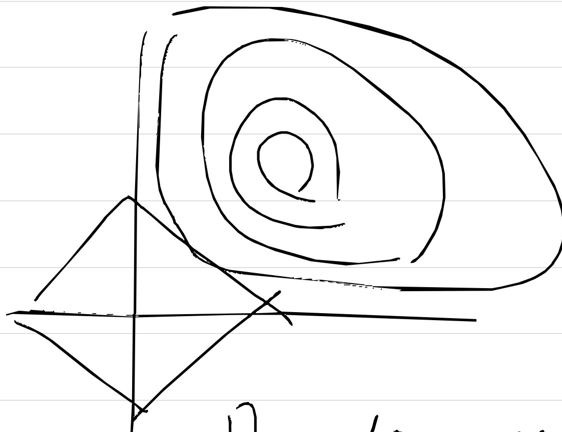
$$x_j'(y - X\hat{\beta}) = \lambda\,\mathrm{Sign}(\hat{\beta}_j)$$

if $\hat{\beta}_j = 0$, then

$$|x_j'(y - X\hat{\beta})| < \lambda$$

a local pertubation of $\hat{\beta}_j$ cannot compensate the penalty brought by λ. Better to stay at $\hat{\beta}_j = 0$

Geometry of lasso



equivalent expression

$$\min \frac{1}{2n} \|y - X\beta\|_2^2$$

$$\text{s.t. } \|\beta\|_1 \le C$$

Depending on the shape of the contour it is likely to find corner solutions.

Lasso is viewed as a variable selector (Tibshirani, 1996)
But lasso does consistently select the true model under very restrictive conditions.
(Zou, 2006)

Determining the tuning parameter

information criterion

$$\text{AIC: } \log \hat{\sigma}^2 + \frac{2}{n} \hat{p}(\lambda)$$

$$\text{BIC: } \log \hat{\sigma}^2 + \frac{\log n}{n} \hat{p}(\lambda)$$

for ridge $\hat{p}(\lambda) = \sum_j \frac{d_j}{d_j + \lambda}$.

data-driven approach.

① sample splitting

② cross validation

Asy. of lasso

if $\lambda = C\sqrt{\frac{\log p}{n}}$ at some speed, then

$$\frac{1}{n}\|X'\hat{\beta} - X'\beta_0\|_2^2 \xrightarrow{p} 0$$

$$\|\hat{\beta} - \beta_0\|_1 \xrightarrow{p} 0$$

$$\|\hat{\beta} - \beta_0\|_2 \xrightarrow{p} 0.$$

Regularization methods for variable selection

SCAD, MCP. Smoothly clipped abs. deviation (2001)
minimax concave penalty

Zhang Cui Hui (2010)

nonconvex optimization

variable selection consistency.

Feature engineering for penalized methods

normalization: demean, and scale-norm.

affect finite sample performance

An overview of regularization methods

* Bias-variance trade-off

Supervised learning.

① training. use $(y, X)_{train}$. find $\hat{f}_\lambda$

② validation find $\hat{\lambda}$

$\Downarrow$

find a trained model $\hat{f}_{\hat{\lambda}}$

③ test data $(y, X)_{test}$.

use $\hat{f}_{\hat{\lambda}}(X_{test})$ to predict y_{test}.

For regression problems, MSEloss is the most popular

$$MSEloss = \frac{1}{n}\| y - \hat{f}(x) \|_2^2$$

Equivalent to

$$\frac{1}{n} \| E(y|X) - f(x) \|_2^2$$

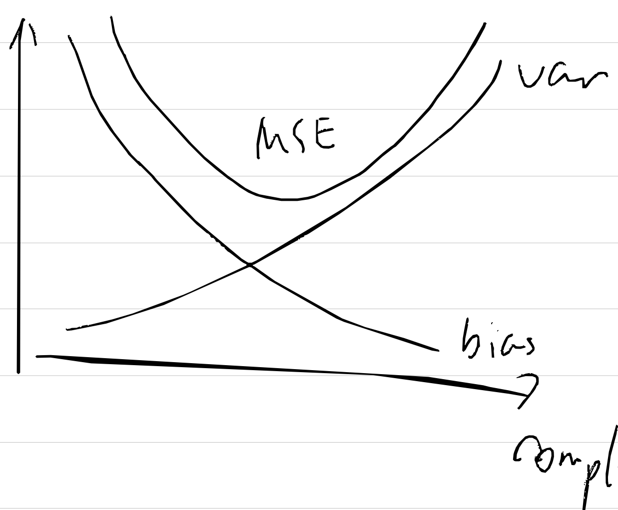
Since $y = E(y|X) + \varepsilon$, where ε is unpredictable

$$\frac{1}{n} \| E(y|X) - E[\hat{f}(x)] + E[\hat{f}(x)] - f(x) \|_2^2$$

$$= \frac{1}{n} \| E(y|X) - E[\hat{f}(x)] \|_2^2 + \frac{1}{n} \| E[\hat{f}(x)] - f(x) \|_2^2$$

$$= \quad \text{bias}^2 \qquad\qquad\qquad \text{var.}$$

$$MSE = \text{bias}^2 + \text{var}$$



look for tuning parameters to reduce MSE

Tuning parameters are also called hyperparameters

it cannot be easily embeded into a criteria for optimization

Classical econometrics generative model

1. population model (identification)
2. estimation (point estimation)
3. inference (interval estimation) eff.
4. interpretation (story telling)

Machine learning focuses on predictive performance in test data

* Some methods do not have DGP
* lack of interpretability

---

econometrics	ML
$y = x'\beta + \varepsilon$	$y = f(x) + \varepsilon$
learn β	learn $f(x)$
no tuning parameter due to assumption	tuning parameter
focus on inference of β	focus on predictibility

My research:

- variable selection: Lee, Shi & Gao (2022)
- feature engineering: Mei & Shi (2024)
- GMM-lasso: Shi (2016)
- forward selection: Shi & Huang (2023)
- ridge-type boosting: Phillips & Shi (2021)