理论与性质：罚项光滑性

$\hat{\beta}=(X^TX)^{-1}X^TY$

若 X^TX 不可逆 $\Rightarrow \hat{\beta}$ 无法计算

Ridge:

$\hat{\beta}(k)=(X^TX+kI)^{-1}X^TY$

不可逆 $\Rightarrow X^TX$ 特征值存在0

$\Rightarrow |X^TX|=0=\prod_{i=1}^{p}\lambda_i$

→半正定 $\Rightarrow \lambda_i\geq 0$

Prop. $\{\lambda_1,\cdots,\lambda_p\}$ 为 X^TX 的特征值

$\Rightarrow \{\lambda_1+k,\lambda_2+k,\cdots,\lambda_p+k\}$ 为 (X^TX+kI) 的特征值

$(k>0)$

Pf. ① X^TX s.p.d. 半正定

Def of s.p.d. $\forall y\in R^{p\times 1}$，有

$\Rightarrow y^T\Sigma y\geq 0 \Rightarrow \Sigma$ s.p.d

Given $y\in R^{p\times 1}$

$\Rightarrow y^TX^TXy=(Xy)^TXy\geq 0$

$\Rightarrow X^TX$ s.p.d □

Recap: $Z^TZ\geq 0,\ \forall Z$

Assume: $Z=(z_1,\cdots,z_p)^T$

$\Rightarrow Z^TZ=(z_1,\cdots,z_p)\begin{pmatrix}z_1\\ \vdots\\ z_p\end{pmatrix}=\sum_{i=1}^{p}z_i^2\geq 0$

Eigenvalue: 若 $x\in R^{p\times 1}$，$\lambda\in R$，且 $Ax=\lambda x$

(特征值) 则 λ 为 A 的某个特征值，x 为 A 的… 对应的特征向量 (Eigenvectors)

Assume: $Ax=\lambda x$

$\Rightarrow (A+kI)x=Ax+kx=(A+k)x=(\lambda+k)x$

$\Rightarrow \lambda+k$ 为 $A+kI$ 的 …，…

若 λ 为 A 的 eigenvalue $\Rightarrow \lambda+k$ 为 $(A+kI)$ 的 eigenvalue

$|X^TX|=0$

$\Rightarrow |X^TX|\approx 0 \Rightarrow \hat{\beta}$ 不稳定 $[D(\hat{\beta})=\sigma^2(X^TX)^{-1}]$

希望β波动别太大 $\Rightarrow$ β加限制

$\beta(k) = (X^TX+kI)^{-1}X^TY$

$\searrow$ 与$\boxed{\text{稳定}}$关联的参数

$f(\beta) = \sum (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

$= (y - X\beta)^T (y - X\beta) + \underbrace{\lambda \beta^T \beta}_{} \rightarrow 2\lambda\beta$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nn} \end{pmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

求导 $\frac{\partial f(\beta)}{\partial \beta} = \nabla f(\beta)$?

令 $g(\beta) = (y - X\beta)^T(y - X\beta)$

$\Rightarrow \lim_{t\to 0} \frac{g(\beta+tv) - g(\beta)}{t}$ $(z = y - X\beta)$

$= \lim_{t\to 0} \frac{(z - tXv)^T(z - tXv) - z^Tz}{t}$

$= \lim_{t\to 0} \frac{-tz^TXv - tv^TX^Tz}{t} = -2v^TX^T(y - X\beta)$

$= v^T \cdot \nabla g(\beta)$

$\because v$ 任意 $\Rightarrow \nabla g(\beta) = -2X^T(y - X\beta)$

$\Rightarrow \nabla f(\beta) = -2X^T(y - X\beta) + 2\lambda\beta = 0$

$\Rightarrow X^Ty = X^TX\beta + \lambda\beta = (X^TX + \lambda I)\beta$

$\Rightarrow \hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$

一阶导为零的点 $\Rightarrow$ 极值

e.g. $f(x) = x^3 \Rightarrow f'(x) = 3x^2 \Rightarrow f'(0) \Rightarrow$ but $x \Rightarrow$ 不是极值

极值 ⇒ 二阶 >0（海塞矩阵正定）

$$\lim_{t\to 0}\frac{\nabla f(\beta+tv)-\nabla f(\beta)}{t}$$

$\nabla f(\beta) = -2(\underbrace{X^T y}_{\text{与}\beta\text{无关}} - X^T X\beta - \lambda\beta)$

$$= \lim_{t\to 0}\frac{2(tX^T X v + \lambda t v)}{t} = 2(X^T X + \lambda I)v$$

$$= \nabla^2 f(\beta)\cdot v$$

∵ v 任意

∴ $\nabla^2 f(\beta) = X^T X + \lambda I > 0$ p.d

含义与解释：奇异值分解

收缩：$\hat\beta(k) = (X'X+kI)^{-1}X'Y = A_k\hat\beta$

其中 $A_k = (X'X+kI)^{-1}X'X$

奇异值分解（SVD）

$\forall X \in R^{n\times m}$

$\Rightarrow X = U\Sigma V^T$ $\quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$

（U, V 是方阵，$U^TU=I$，$V^TV=I$，Σ 为对角阵）

不起作用

$=0$

$$\Rightarrow \underbrace{\begin{bmatrix} u_1 & u_2 & \cdots & u_{p+1} \cdots u_m \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_{p+1} & \\ & & & & \sigma_m \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_m^T \end{bmatrix}}_{V^T}$$

若 $\sigma_m = 0 \Rightarrow u_m, v_m^T$ 不起作用

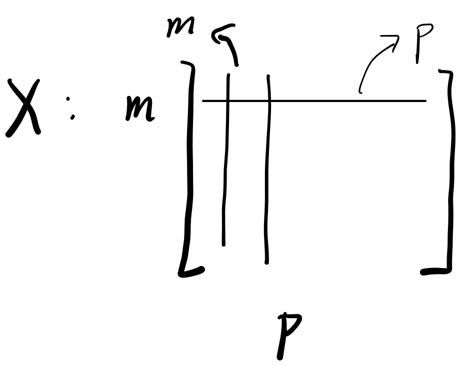
若 $\sigma_{p+1}, \cdots, \sigma_m = 0 \Rightarrow u_{p+1}, \cdots, u_m, v_{p+1}^T, \cdots, v_m^T$ 均不起作用

$$\Rightarrow \underset{m\times p}{[u_1\ u_2 \cdots u_p]} \begin{bmatrix} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_p & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix} \underset{p\times m}{\begin{bmatrix} v_1^T \\ \vdots \\ v_p^T \end{bmatrix}}$$

非方阵

$$\xrightarrow[\text{修改}]{} \underset{m\times p}{[u_1\ u_2 \cdots u_p]} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \\ & & & \sigma_p \end{bmatrix} \underset{p\times p}{\begin{bmatrix} v_1^T \\ \vdots \\ v_p^T \end{bmatrix}}$$

U 列信息　　Σ 奇异值　　V 行信息



$\|\cdot\| \rightarrow$ 范数 (norm)

$\|x\|_2$, $x = (x_1, x_2, \cdots, x_n)^T$

$\Rightarrow \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

定理：对 $\forall k > 0$, $\|\beta(k)\| < \|\beta\|$

证：设 $\underset{m\times p}{X}=\underset{m\times p}{U}\underset{p\times p}{\Sigma}\underset{p\times p}{V^T}$

$\Rightarrow X^TX=V\Sigma \underline{U^TU}\Sigma V^T$
$=V\Sigma^2V^T$

$\Rightarrow X^TX+kI=V(\Sigma+kI)V^T$

$[V^TV=I$，V是方阵 $\Rightarrow V^T=V^{-1}\Rightarrow VV^T=I)$

$\Rightarrow (X^TX+kI)^{-1}=V(\Sigma^2+kI)^{-1}V^T$

$\Rightarrow (X^TX+kI)^{-1}X^TX=V(\Sigma^2+kI)^{-1}V^T\cdot V\cdot\Sigma^2V^T$
$=V(\Sigma^2+kI)^{-1}\Sigma^2\cdot V^T$

又 $\hat{\beta}(k)=A_k\cdot\hat{\beta}$

$\Rightarrow \|\hat{\beta}(k)\|=\|V(\Sigma^2+kI)^{-1}\Sigma^2\cdot V^T\hat{\beta}\|$

$\|AB\|\le\|A\|\cdot\|B\|$

$\le\|V\cdots V^T\|\cdot\|\hat{\beta}\|$

$\|UX\|=\|X\|$，U列正交

$\|XU\|=\|X\|$，U行正交

$\Rightarrow \|X\|=\|\Sigma\|$，$X=U\Sigma V^T$

$=\|(\Sigma^2+\lambda I)^{-1}\cdot\Sigma^2\|\|\hat{\beta}\|<\|\hat{\beta}\|$

$$\begin{bmatrix}\frac{1}{\sigma_1^2+\lambda} & & & \\ & \frac{1}{\sigma_2^2+\lambda} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_p^2+\lambda}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$$

$$=\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2+\lambda} & & \\ & & \ddots & \\ & & & \frac{\sigma_p^2}{\sigma_p^2+\lambda} \end{bmatrix}$$

$\Rightarrow \|\sim\| \le \|I\| = 1$ 矩阵的二范数

由 $f(x) = \frac{x^2}{x^2+\lambda} = 1 - \frac{\lambda}{x^2+\lambda}$，$f(x)$递增

$\Rightarrow \sigma_i^2 \uparrow$，$f(\sigma_i) \uparrow$，越不容易被收缩

SVD的几何意义

e.g. $X = U\Sigma V^T$

$$\underset{(2\times 2)}{} = [u_1 \ u_2] \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$$

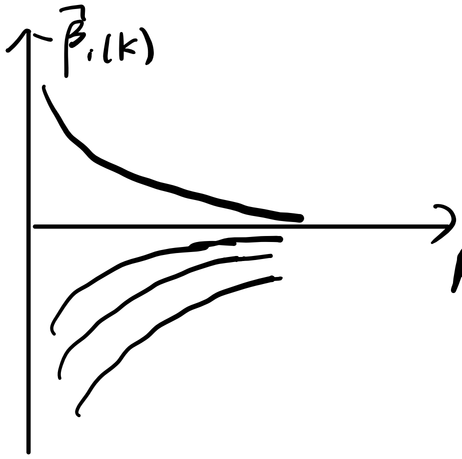
$\Rightarrow \sigma_i \uparrow$，信息越多，越不易收缩

$\Rightarrow$ 岭回归 保留信息多的，剔除信息少的

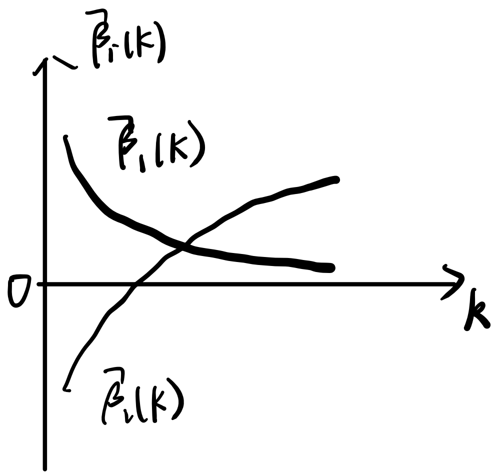
$\Rightarrow u_1 \rightarrow$ 信息多的方向

$u_2 \rightarrow \sim$ 第二多的方向

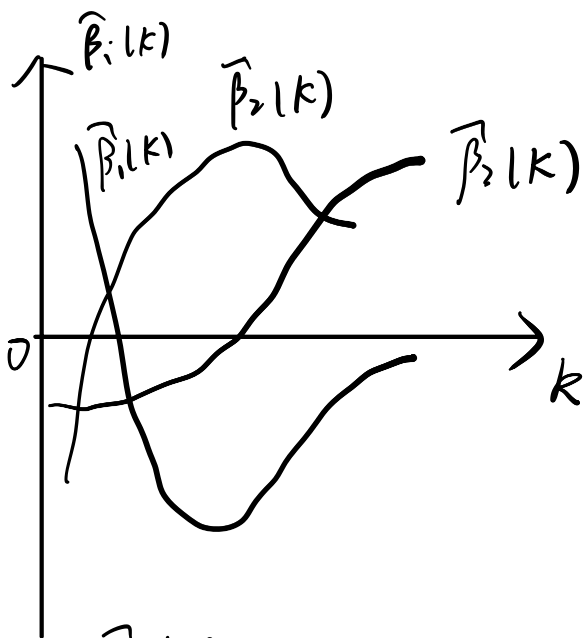
岭迹分析



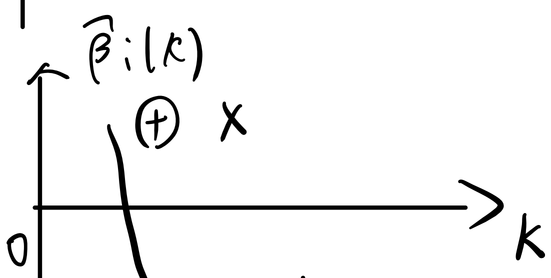
← 稳定
不变号
趋于0
⇒ 健康
稳定(原始最小二乘模型可使用)



变号
一升一降(负相关) ⇒共线性
⇒具有共线性的问题
Solution: 删掉一个变量



原模型不稳定

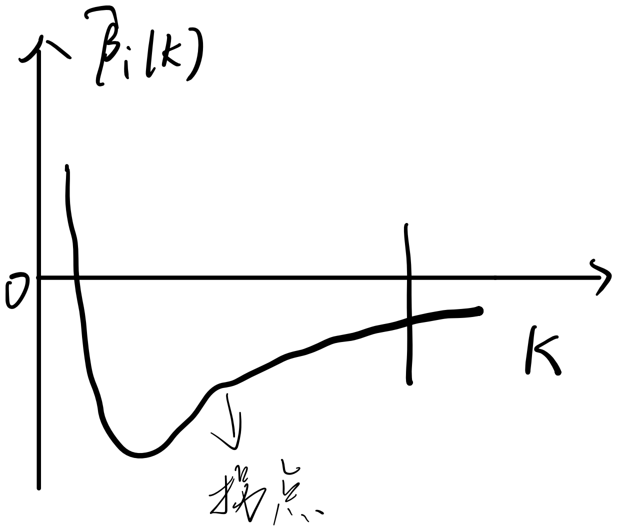


变号

$\beta>0$ ⇒ 正相关

$<0 \Rightarrow$ 负相关

$\Rightarrow$ 看似正相关,实则负相关



此变量无作用

($\beta \approx 0$)

LASSO

岭回归:

Definition 4: Ridge Regression Estimator

We denote $\boxed{\hat{\beta}(k) = (\mathbf{X}^T\mathbf{X} + kI)^{-1}\mathbf{X}^T\mathbf{Y}}$ the Ridge Regression Estimator.

$\hat{y}_i$

罚项

Ridge Regression Objective

$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \boxed{\beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j})^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

Obviously, $\lambda \sum_{j=1}^{p} \beta_j^2$ is the penalty. (Meaning?)

LASSO:

LASSO Objective

$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

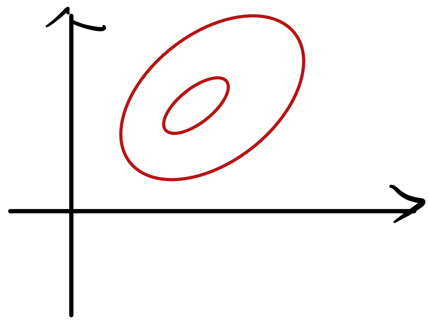
罚项

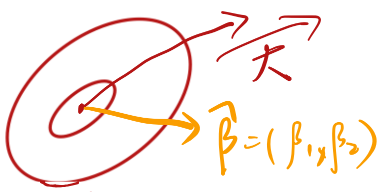
$$f(x;\beta)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

对偶性

$$\Rightarrow \min \boxed{(y_i-\hat{y}_i)^2} \quad 二次(凸)$$

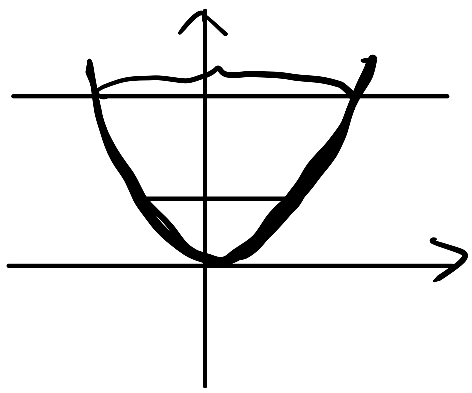
$$\text{s.t.} \sum_{j=1}^{p}|\beta_j| \le t$$





圆越来越大
取值越来越大

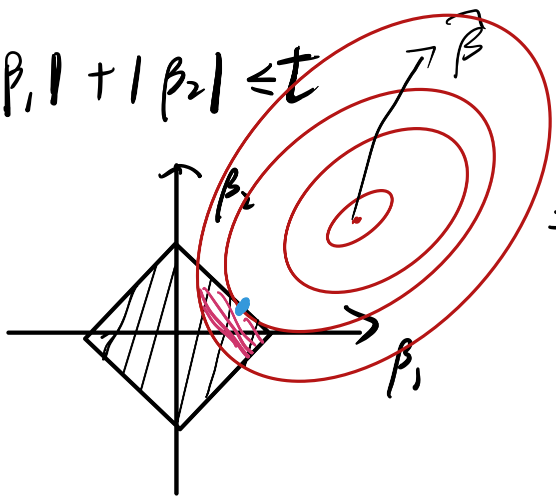
$\Rightarrow$



↓ view

等高线图
二维视角

$|\beta_1|+|\beta_2|\le t$



$\lambda=0$

$\Rightarrow t=\infty$

约束条件：area尽可能小