回顾：多重共线性，即 $\exists\ \lambda_{min}(X^TX)$ 非常接近于 0

一、岭回归定义

Definition of Ridge Regression

1.1 调节参数平移

记 $k>0$ 为调节参数（tuning parameter），我们把设计矩阵 X^TX 平移 k 个单位后进行 OLSE，于是得到 β 的岭回归估计

$$\hat{\beta}_k = (X^TX + k\cdot I)^{-1}\cdot X^T\cdot y$$

显然，原始 OLSE 是 $k=0$ 时的特例：$\hat{\beta} = \hat{\beta}_0$

$$\boxed{\text{Lemma: } \lambda_{min}(X^TX + k\cdot I) \geq k}$$

Proof：$\because$ X^TX 为实对称阵 $\Rightarrow$ $X^TX = \Phi\Lambda\Phi^T$ 其中 Φ 为正交阵，而 Λ 为对角阵。则有

$$X^TX + kI = \Phi\Lambda\Phi^T + kI = \Phi(\Lambda + kI)\Phi^T$$

于是有 $\lambda(X^TX + k\cdot I)$ 为 $\Lambda + kI$ 的对角元素

又注意到 $v^TX^TXv = \|Xv\|^2 \geq 0$，则有 X^TX 半正定，Λ 的每个元素非负

故 $\forall\ \lambda(X^TX + k\cdot I) \geq 0 + k = k$ $\square$

由 Lamma，我们可以保证 $(X^TX + kI)$ 避免多重共线性。

1.2 基于优化的定义

岭回归估计 $\hat{\beta}_k$ 也可以通过如下定义：

$$\hat{\beta}_k = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - Xb\|^2 + k\|b\|^2 \right\}$$

本质为"惩罚回归"(penalized regression)，且采用了 L_2 范数惩罚项。

记 $Q(b) = \|y-Xb\|^2 + k\|b\|^2$，则有

$$\frac{\partial Q}{\partial b} = 2X^T(y-Xb) + 2kb$$

$$\frac{\partial^2 Q}{\partial b^2} = 2X^TX + 2kI$$

令一阶导为0 得 $\hat{\beta}_k = (X^TX + kI)^{-1}X^Ty$ 与定义 1.1 等价

Proof: 由 Lemma 知 X^TX+kI 的特征根 $\geq k > 0$，故 X^TX+kI 正定，即 $\frac{\partial^2 Q}{\partial b^2} \geq 0$ 得 $\hat{\beta}_k$ 确为最小值点。

1.3 基于 Lagrange 定义

岭回归也可以定义为有限制的优化问题：$(s \geq 0)$

$$\tilde{\beta}_s = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \|y - Xb\|^2$$

$$\text{s.t.} \quad \|b\|^2 \leq s$$

如上同样采用了 L_2 范数约束，下面证明等价性：

Proof: 对 $\forall s\in[0,\infty]$，都 $\exists k\in[0,\infty]$，使 $\hat{\beta}_k=\tilde{\beta}_s$

(1) 若 $\mathrm{rank}(X) < p$ OLSE 不存在，我们不讨论

(2) 若 $\mathrm{rank}(X) = p$

① 若 $(X^TX)^{-1}X^Ty = 0$，即 OLSE: $\hat{\beta} = 0$ 那么对 $\forall s$ $\|b\|^2 \le s$ 无意义

此时 $\hat{\beta}_k = \tilde{\beta}_s = 0$ 平凡

② 若 $(X^TX)^{-1}X^Ty \ne 0$，由 Lagrange 乘子法：

$$L(b,u) = \|y-Xb\|^2 + u\cdot(s-\|b\|^2)$$

$$||y-Xb||^2 + u\cdot(||b||^2 - s) = 0$$

故求导知，$\tilde{\beta}_s$ 满足：

$$-2X^T(y-X\tilde{\beta}_s)+u\tilde{\beta}_s=0 \qquad u\cdot(\|\tilde{\beta}_s\|^2 - s) = 0 \quad (1)$$

$$||\tilde{\beta}_s||^2 - s = 0 \qquad X^T(X\tilde{\beta}_s - y) + u\cdot\tilde{\beta}_s = 0 \quad (2)$$

(a) 若 $s \ge \|\hat{\beta}\|^2 = \|(X^TX)^{-1}X^Ty\|^2$，则限制无意义

取 $k=0$ 有 $\hat{\beta}_k=\hat{\beta}=\tilde{\beta}_s$

(b) 若 $0\le s < \|(X^TX)^{-1}X^Ty\|^2$

则 $u\ne0$，否则由(2)有 $\|\tilde{\beta}_s\|^2 = \|(X^TX)^{-1}X^Ty\|^2 > s$ 矛盾！

由(2)知

$$\tilde{\beta}_s = (X^TX+u\cdot I)^{-1}X^Ty \quad (3)$$

由(1)知

$$s = \|\tilde{\beta}_s\| = \|(X^TX+u\cdot I)^{-1}X^Ty\|^2 \quad (4)$$

故对给定的 s，由(4)解出 u，令 $k=u$ 有 $\hat{\beta}_k=\tilde{\beta}_s$ □

(∵ $f(u)=\|(X^TX+u\cdot I)^{-1}X^Ty\|^2$ 连续，且值域与 s 取值重合，一定有解 u)

1.4 基于 Bayesian 学派的定义

假设似然函数为

$$y|\beta \sim N(X\beta, I),$$

先验 (prior) 分布为

$$\beta \sim N(0, \tau^2 I)$$

其中 $\tau > 0$，则后验 (posterior) 分布为

$$P(\beta|y) = \frac{P(\beta)\cdot P(y|\beta)}{P(y)}$$

$$\propto P(\beta)\cdot P(y|\beta)$$

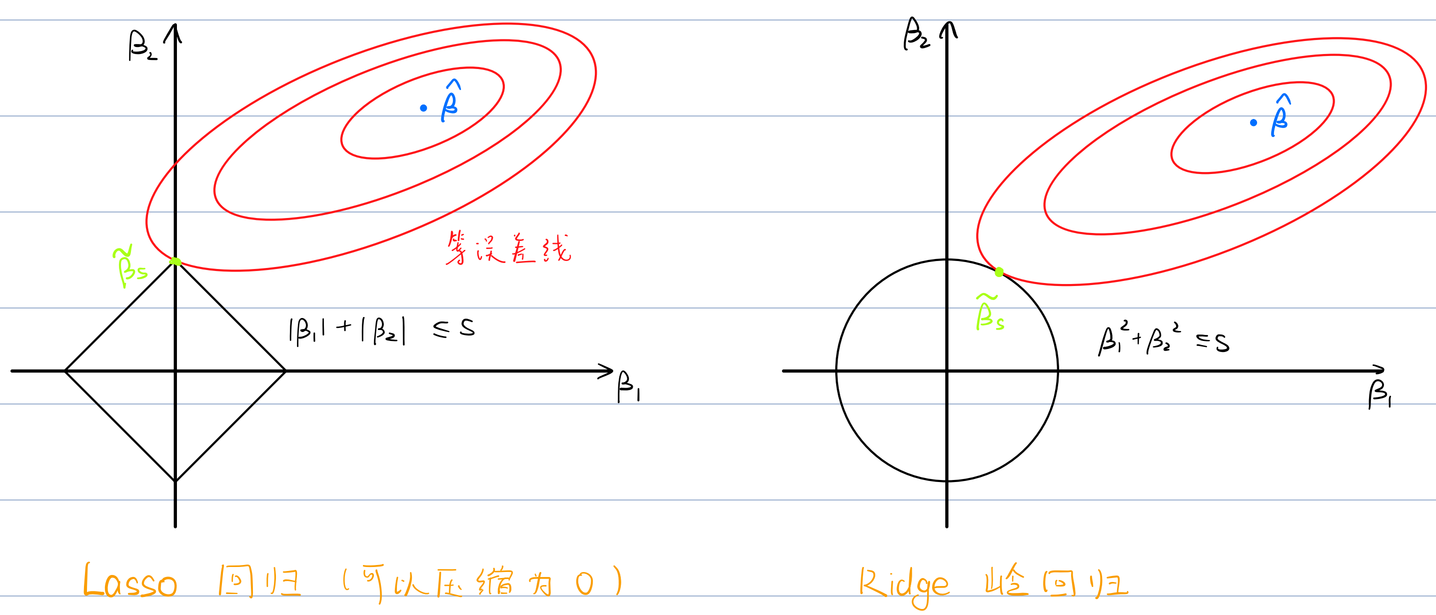
$$\propto \exp\left\{-\frac{\|\beta\|^2}{2\tau^2}\right\}\cdot \exp\left\{-\frac{1}{2}\|y-X\beta\|^2\right\}$$

$$= \exp\left\{-(2\tau^2)^{-1}\cdot\|\beta\|^2 - \frac{1}{2}\|y-X\beta\|^2\right\}$$

故 β 的最大后验估计为

$$\hat{\beta}_\tau = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\|y-Xb\|^2 + \tau^{-2}\|b\|^2\right\}$$

等价于之前定义，取 $k = \tau^{-2}$ 即可



二、岭回归性质

Properities of Ridge Regression

2.1 线性 Linearity

若认为调节参数 k 与 y 无关，则

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

是 y 的线性函数，

$$\begin{aligned}\hat{\beta}_k &= (X^TX + kI)^{-1}X^Ty \\ &= (X^TX + kI)^{-1}\cdot X^TX(X^TX)^{-1}\cdot X^Ty \\ &= (X^TX + kI)^{-1}X^T\cdot X\cdot (X^TX)^{-1}X^Ty \\ &= (X^TX + kI)^{-1}X^TX\hat{\beta}\end{aligned}$$

是 $\hat{\beta}$ 的线性函数

Tips：实际 k 取值依赖于 y，上面就不成立

2.2 有偏性 biasedness

$$E(\hat{\beta}_k) = (X^TX + kI)^{-1}\cdot X^TX\beta$$

若 $k>0$，则 $E(\hat{\beta}_k) \neq \beta$

2.3 估计值的 L_2 范数减小 Shrinkage

若 $\hat{\beta} \neq 0$，$k > 0$ 则有

$$\|\hat{\beta}_k\| < \|\hat{\beta}\|$$

Tips：但无法压缩至以正概率取 0，但 LASSO 回归可以：

利用 L_1 范数修正，可以使 $\exists j$ $P(\hat{\beta}_{k[j]} = 0) > 0$

2.4 协方差减小 Shrinkage of Covariance

若 $k > 0$，则 $\{cov(\hat{\beta}) - cov(\hat{\beta}_k)\}$ 正定

Proof：见习题

Tips：性质 2.3 和 2.4 说明，岭回归对估计量进行了压缩（Shrinkage）

显然，当 $k \to \infty$，$\hat{\beta}_k \to 0$

由 Chapter 4 Theorem 5：

$$\begin{aligned} E(\|\hat{\beta}\|^2) &= E(\hat{\beta}^T I \hat{\beta}) \\ &= E(\hat{\beta}^T) \cdot I \cdot E(\hat{\beta}) + tr\{I \cdot cov(\hat{\beta})\} \\ &= \|\beta\|^2 + \sigma^2 \sum_{i=1}^{p} \lambda_j^{-1} \end{aligned}$$

其中 $\lambda_1, \cdots, \lambda_p$ 为 X^TX 的特征值。这说明，$\lambda_{min} \to 0$ 时，

OLSE 的 L_2 范数期望 $E(\|\hat{\beta}\|^2) \to \infty$.

2.5 均方误差 mean squared error, MSE

对于 β 的任意估计量 $\tilde{\beta}$，定义其均方误差为：

$$MSE(\tilde{\beta}) = E(\|\tilde{\beta}-\beta\|^2)$$

则岭回归对均方误差的改进：

$$\exists\, k \quad s.t. \quad MSE(\hat{\beta}_k) < MSE(\hat{\beta})$$

即： $$E(\|\hat{\beta}_k-\beta\|^2) < E(\|\hat{\beta}-\beta\|^2)$$

Lemma 1 对于 β 的任意估计量 $\tilde{\beta}$，其 MSE 均满足：

$$MSE(\tilde{\beta}) = tr\{cov(\tilde{\beta})\} + \|E(\tilde{\beta})-\beta\|^2$$

此外，对于任意正交阵 $P \in \mathbb{R}^{p\times p}$，有：

$$E\{\|P\cdot\tilde{\beta}-P\beta\|^2\} = E(\|\tilde{\beta}-\beta\|^2)$$

利用 Lemma 1 对性质 2.5 证明：

由于 X^TX 为实对称矩阵，故 $\exists$ 正交阵 $\Phi \in \mathbb{R}^{p\times p}$ 与对角阵 $\Lambda = diag(\lambda_i)$

其中 λ_i 为 X^TX 的特征根，有：

$$X^TX = \Phi\Lambda\Phi^T$$

记 $Z = X\Phi \quad \alpha = \Phi^T\beta := (\alpha_{[1]}, \cdots, \alpha_{[p]})^T \quad \hat{\alpha}_k = \Phi^T\hat{\beta}_k$，

故由引理第2个恒等式，只须证：

$$E(\|\hat{\alpha}_k-\alpha\|^2) < E(\|\hat{\alpha}_0-\alpha\|^2)$$

分别计算 $cov(\hat{\alpha}_R)$, $E(\hat{\alpha}_R)$ 有：

$$cov(\hat{\alpha}_R) = cov(\Phi^T \hat{\beta}_R)$$

$$= \Phi^T \cdot cov((X^TX+kI)^{-1}X^Ty) \cdot \Phi$$

$$= \Phi^T \cdot (X^TX+kI)^{-1}X^T \cdot cov(y) \cdot X(X^TX+kI)^{-1}\Phi$$

$$= \sigma^2 \cdot \Phi^T \cdot \{\Phi(\Lambda+kI)\Phi^T\}^{-1} X^TX \{\Phi(\Lambda+kI)\Phi^T\}^{-1}\Phi$$

$$= \sigma^2 \cdot \Phi^T \cdot \Phi(\Lambda+kI)^{-1}\Phi^T X^TX \Phi(\Lambda+kI)^{-1}\Phi^T\Phi$$

$$= \sigma^2 (\Lambda+kI)^{-1}\Lambda(\Lambda+kI)^{-1} \quad \text{均为对角阵}$$

$$= \sigma^2 \,\mathrm{diag}\left(\frac{\lambda_i}{(\lambda_i+k)^2}\right)$$

$$E(\hat{\alpha}_R) = E(\Phi^T\hat{\beta}_R)$$

$$= \Phi^T(X^TX+kI)^{-1}X^T E(y)$$

$$= \Phi^T(X^TX+kI)^{-1}X^T X\beta$$

$$= \Phi^T\Phi(\Lambda+kI)^{-1}\Phi^T \cdot (\Phi\Lambda\Phi^T)\cdot\beta$$

$$= (\Lambda+kI)^{-1}\Lambda\Phi^T\beta$$

$$= (\Lambda+kI)^{-1}\Lambda\alpha \quad \text{均为对角阵}$$

$$= \mathrm{diag}\left(\frac{\lambda_i}{\lambda_i+k}\right)\cdot\alpha$$

由引理可得 $MSE(\hat{\alpha}_R)$ 为

$$g(k) = MSE(\hat{\alpha}_R) = E(\|\hat{\alpha}_R-\alpha\|^2)$$

$$= tr(cov(\hat{\alpha}_R)) + \|E(\hat{\alpha}_R)-\alpha\|^2$$

$$= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2} + k^2\sum_{i=1}^{p}\frac{\alpha_{[i]}^2}{(\lambda_i+k)^2} \quad \in \mathbb{R}$$

求导得：

$$\frac{dg}{dk} = -2\sigma^2\cdot\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3} + 2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_{[i]}^2}{(\lambda_i+k)^3}$$

$\because\ g'(0) = -2\sigma^2\sum\frac{1}{\lambda_i^2} + 0 < 0$ 而 $g'(k)$ 连续，故 $\exists\, k_0 > 0$

使 $g'(k) < 0$ 对 $\forall k \in (0, k_0]$ 成立，于是：

$$\text{对 } \forall k \in (0, k_0] \text{ 有 } g(k) < g(0)$$

$$\text{即 } E(\|\hat{\alpha}_k - \alpha\|^2) < E(\|\hat{\alpha}_0 - \alpha\|^2) \qquad \blacksquare$$

三、调节参数 k 的选取

Choice of the tuning parameter k

3.1 交叉验证 Cross Validation

对固定 $M \in N_+$，$M \geq 2$. 将指标集合 $J = \{1, 2, \cdots, n\}$ 分割为 M 个容量相近的互斥子集 $J_1, \cdots, J_M$.

对于 $m = 1, 2, \cdots, M$，记 $\hat{\beta}_k^{(-m)}$ 为在 $\{(y_i, x_i) : i \notin J_m\}$ 上训练的岭回归估计值。

针对 k 的选取集 $K \subset [0, \infty)$，取 k 为使如上表现最佳的：

$$\hat{k} = \operatorname*{argmin}_{k \in K} \sum_{m=1}^{M} \sum_{i \in J_m} (y_i - x_i^T \hat{\beta}_k^{(-m)})^2$$

Tip：Cross Validation 是最常用的方法

3.2 Hoerl-Kennad 公式

3.3 Mcdorand-Gararneau 法

3.4 双h公式

见讲义