岭回归（Ridge Regression）

TOC

基本模型

- **线性回归模型**：

$$Y = X\beta + \varepsilon$$

- **普通最小二乘（OLS）估计**：

$$b = (X^T X)^{-1} X^T Y$$

- **岭回归估计**：

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

特殊情形：$X^T X = I$

当 $X^T X = I$（即 X 为正交矩阵）时：

- 岭回归估计简化为：

$$\hat{\beta}(\lambda) = \frac{1}{1+\lambda} X^T Y$$

- 而 OLS 估计为：

$$b = X^T Y$$

- 进一步得：

$$\hat{\beta}(\lambda) = \frac{1}{1+\lambda} b$$

极限情况：

- 当 $\lambda = 0$，$\hat{\beta}(0) = b$（退化为 OLS）；
- 当 $\lambda \to \infty$，$\hat{\beta}(\infty) = 0$。

一般情形：$X^T X$ 非单位矩阵的分析

矩阵分解与逆矩阵形式

对 $X^T X$ 进行特征值分解：

$$X^T X = U D U^*$$

则：

$$(X^T X + \lambda I)^{-1} = U (D + \lambda I)^{-1} U^*$$

其中，原特征根 d_i 被压缩为：

$$\frac{d_i}{d_i + \lambda}$$

岭回归的偏差与方差（均方误差分析）

期望（偏差分析）

岭回归估计的期望：

$$E[\hat{\beta}(\lambda)] = (X^T X + \lambda I)^{-1} X^T X \beta$$

偏差为：

$$\text{bias} = E[\hat{\beta}(\lambda)] - \beta = -\lambda (X^T X + \lambda I)^{-1} \beta$$

协方差（方差分析）

岭回归估计量的协方差矩阵：

$$\text{Cov}[\hat{\beta}(\lambda)] = \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$$

且满足：

$$\text{Cov}[\hat{\beta}(\lambda)] \leq \text{Cov}(b) = \sigma^2 (X^T X)^{-1}$$

即：**岭回归方差小于 OLS 方差。**

均方误差（MSE）

MSE 分解为：

$$\begin{aligned}
\text{MSE} &= E\|\hat{\beta}(\lambda) - \beta\|^2 \\
&= \text{Var}[\hat{\beta}(\lambda)] + \|\text{bias}\|^2 \\
&= \sigma^2 \text{tr}\left[X^T X (X^T X + \lambda I)^{-2}\right] + \lambda^2 \beta^T (X^T X + \lambda I)^{-2} \beta
\end{aligned}$$

特殊情形（$X^T X = I_{k+1}$）：

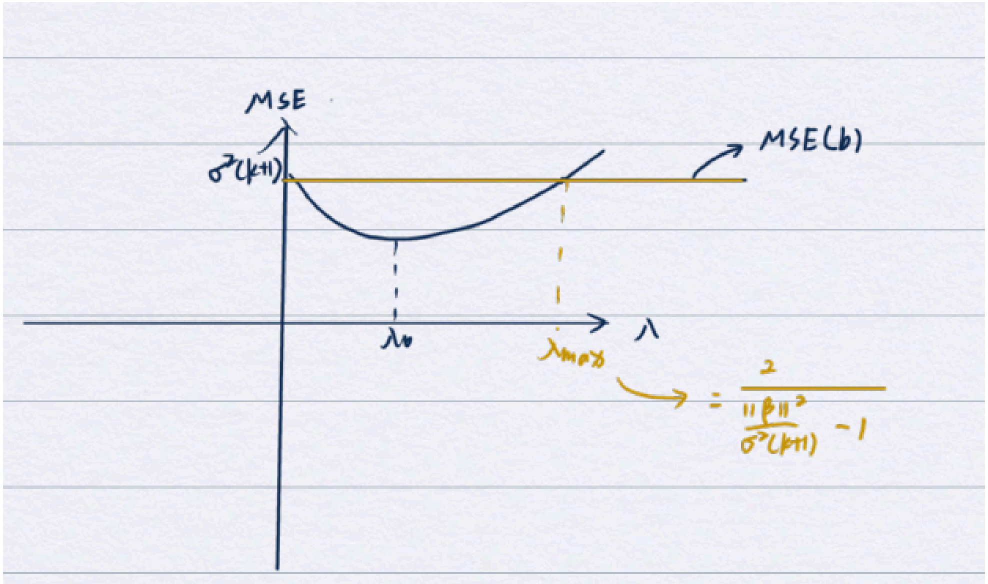
$$\text{MSE}(\hat{\beta}(\lambda)) = \frac{\sigma^2 (k+1) + \lambda^2 \|\beta\|^2}{(1+\lambda)^2}$$

令导数为零可得最优岭参数：

$$\lambda_0 = \frac{\sigma^2(k+1)}{\|\beta\|^2}$$

另一个参考估计值：

$$\lambda_{\text{Hoerl}} = \frac{2}{\frac{\|\beta\|^2}{\sigma^2(k+1)} - 1}$$



信号与噪声（SNR）

模型：

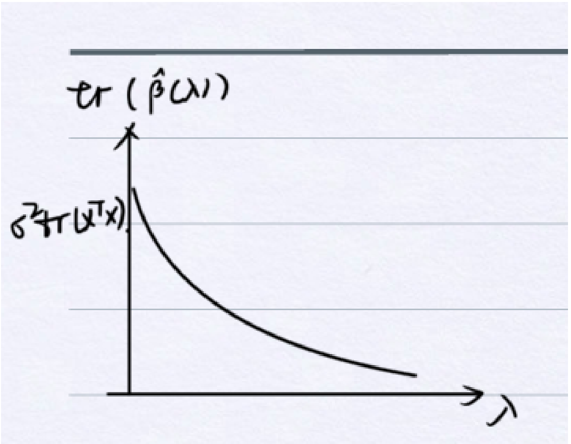
$$Y = X\beta + \varepsilon$$

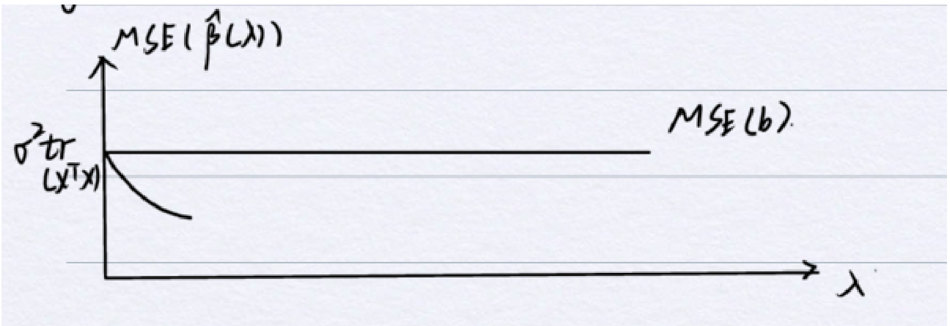
信噪比定义为：

$$\text{SNR} = \frac{\|\beta\|^2/(k+1)}{\sigma^2}$$

- 若 SNR ↑（信号强、噪声弱）→ 倾向于使用 OLS；
- 若 SNR ↓（信号弱、噪声强）→ 倾向于使用岭回归。

图形辅助理解





- **岭估计范数变化**:
  $\|\hat{\beta}(\lambda)\|$ 随 λ 增大单调下降（从 $\|b\|$ 逐渐收缩至 0）。
- **MSE 曲线**:
  MSE 曲线存在最小值点 λ_0，且:

  $$\mathrm{MSE}(\hat{\beta}(\lambda_0)) < \mathrm{MSE}(b)$$

  表明岭回归通过合理选取 λ，能在**偏差—方差权衡**中获得更优的均方误差表现。

$\lambda = 0$ 时 MSE 导数的分析

要分析 $\lambda = 0$ 时 $MSE(\hat{\beta}(\lambda))$ 的导数:

- 首先，MSE 中与方差相关的迹项为 $\sigma^2 \operatorname{tr}(X^TX(X^TX + \lambda I)^{-2})$。
- 对 X^TX 进行奇异值分解 $X^TX = UDU^T$，代入得:

$$
\sigma^2 \operatorname{tr}(UDU^T \cdot U(D+\lambda I)^{-2}U^T) = \sigma^2 \sum_i \frac{d_i}{(d_i+\lambda)^2}
$$

- 对 λ 求导:

$$
\frac{d}{d\lambda} MSE = \sigma^2 \sum_i \frac{-2d_i}{(d_i+\lambda)^3}
$$

- 当 $\lambda = 0$ 时，导数为 $-2\sigma^2 \sum_i \frac{1}{d_i^2} \leq 0$，说明 $\lambda = 0$ 时 MSE 的导数小于 0。

岭回归的预测与核方法关联

岭回归的预测式

已知岭回归系数估计为:

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

则预测值为:

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

矩阵等式推导

可推导得:

$$
(X^T X + \lambda I)^{-1} X^T = X^T (X X^T + \lambda I)^{-1}
$$

其等价性可通过等式 $(X^T X + \lambda I) X^T = X^T (X X^T + \lambda I)$ 验证。

核岭回归的引出

令 $K = XX^T$，则：

$$K[i,j] = (XX^T)[i,j] = x_i^T x_j = \langle x_i, x_j \rangle$$

即向量 x_i 与 x_j 的内积，可衡量二者距离 $d(x_i, x_j)$。

此时预测式可表示为：

$$\hat{Y} = K(K + \lambda I)^{-1} Y$$

这一形式即为 **核岭回归（Kernel Ridge Regression）**。

核函数示例

- 高斯核：$d(x_i, x_j) = \exp\{-\alpha \|x_i - x_j\|^2\}$
- 多项式核：$d(x_i, x_j) = (a + \langle x_i, x_j \rangle)^d$，其中 a, d 为超参数。

岭回归的罚函数视角

岭回归也可理解为带罚函数的优化问题，其估计式是以下优化问题的解：

$$\hat{\beta}(\lambda) = \arg\min_{\hat{\beta}} \|Y - X\hat{\beta}\|^2 + \lambda \|\hat{\beta}\|^2$$

其中，$\lambda \|\hat{\beta}\|^2$ 为 **罚函数（正则项）**，用于惩罚系数的大小，从而缓解多重共线性。

岭回归中岭参数 λ 的选择方法：交叉验证法

选择岭参数 λ 的核心思路是 **交叉验证（Cross Validation）**，具体分为以下三种：

1. 留一法（Leave-One-Out）
2. k-折法（k-Fold）
3. 广义交叉验证（Generalized Cross Validation）

留一法（Leave-One-Out）

方法逻辑

将第 i 个样本拿出作为测试集，其余 $n-1$ 个样本作为训练集，重复 n 次。

具体步骤

- **训练集构造**：$X_{(i)}$ 表示删除第 i 行的矩阵。
- **岭回归估计**：

$$\hat{\beta}_{(i)}(\lambda) = \left(X_{(i)}^T X_{(i)} + \lambda I\right)^{-1} X_{(i)}^T y_{(i)} = \left(\sum_{j\neq i} x_j x_j^T + \lambda I\right)^{-1} \left(\sum_{j\neq i} x_j y_j\right)$$

- **测试集评估**：用第 i 个样本的特征 x_i 和估计系数 $\hat{\beta}_{(i)}(\lambda)$ 计算预测误差 $(y_i - x_i^T \hat{\beta}_{(i)}(\lambda))^2$。
- **交叉验证指标（CV_1）**：

$$CV_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T \hat{\beta}_{(i)}(\lambda))^2$$

- **最优 λ 选择**：找到使 CV_1 最小的 λ。

---

k-折法（k-Fold）

方法逻辑

将样本集分成 k 个不相交的子集 $J_1, J_2, \ldots, J_k$（满足 $|J_k| = \frac{n}{k}$ 且 $J_i \cap J_j = \emptyset$），每次取其中一个子集作为测试集，其余 $k-1$ 个子集作为训练集，重复 k 次。

具体步骤

- **训练集与测试集构造**：第 t 次迭代时，训练集为 $\bigcup_{s\neq t} J_s$，测试集为 J_t。
- **岭回归估计**：

$$\hat{\beta}_{(t)}(\lambda) = \left(\sum_{j\notin J_t} x_j x_j^T + \lambda I\right)^{-1} \left(\sum_{j\notin J_t} x_j y_j\right)$$

- **测试集评估**：对测试集 J_t 中的每个样本 (x_ℓ, y_ℓ)，计算预测误差 $(y_\ell - x_\ell^T \hat{\beta}_{(t)}(\lambda))^2$。
- **交叉验证指标（CV_k）**：

$$CV_k = \frac{1}{k}\sum_{t=1}^{k} \frac{1}{|J_t|} \sum_{\ell\in J_t} (y_\ell - x_\ell^T \hat{\beta}_{(t)}(\lambda))^2$$

- **最优 λ 选择**：找到使 CV_k 最小的 λ。

留一法与 k-折法的关系

当 $k = n$ 时，k-折法就退化为 **留一法**（每个子集仅含一个样本）。

广义交叉验证（Generalized CV）

交叉验证的核心是评估模型在“留一法”（Leave-One-Out, LO O）下的预测误差，广义交叉验证对其进行了简化推导。

- 留一法交叉验证的损失函数：

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}_i(\lambda)\right)^2$$

  其中 $\hat{\beta}_i(\lambda)$ 是**剔除第 i 个样本后**，通过正则化（罚项为 λ）得到的回归系数。

- 广义交叉验证的简化：
  引入“帽子矩阵” $H(\lambda)$，其元素 $H_{ii}(\lambda) \triangleq x_i^T (X^T X + \lambda I)^{-1} x_i$，且矩阵的迹 $\mathrm{tr}H = \sum_{i=1}^{n} H_{ii}(\lambda)$。
  此时可近似 $H_{ii}(\lambda) \approx \frac{\mathrm{tr}H}{n}$，从而广义交叉验证的损失函数简化为：

$$\lambda = \arg\min_{\lambda} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - \frac{\mathrm{tr}H}{n}}\right)^2$$

关键等式的证明：$y_i - x_i^T \hat{\beta}_i(\lambda) = \frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - H_{ii}(\lambda)}$

要推导广义交叉验证，需先证明“留一法预测残差”与“全样本预测残差”的关系，步骤如下：

定义符号与矩阵分解

- 全样本设计矩阵 $X = [x_1, x_2, \ldots, x_n]^T$（$x_i$ 为第 i 个样本的特征向量），响应向量 $y = [y_1, y_2, \ldots, y_n]^T$。
- 剔除第 i 个样本后的设计矩阵 $X_{(i)}$（去掉第 i 行），响应向量 $y_{(i)}$（去掉第 i 个元素）。
- 矩阵恒等式：$X^T X = \sum_{j=1}^n x_j x_j^T$，$X_{(i)}^T X_{(i)} = \sum_{\substack{j=1 \\ j \neq i}}^n x_j x_j^T$，因此 $X^T X - X_{(i)}^T X_{(i)} = x_i x_i^T$。

proof

已知矩阵求逆引理：若 $A^{-1} - B^{-1}$ 存在，则 $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$。

令 $A = X_{(i)}^T X_{(i)} + \lambda I$，$B = X^T X + \lambda I$，则 $B - A = x_i x_i^T$，代入引理得：

$$A^{-1} - B^{-1} = A^{-1} x_i x_i^T B^{-1}$$

对等式两边**右乘** x_i，得：

$$A^{-1} x_i - B^{-1} x_i = A^{-1} x_i x_i^T B^{-1} x_i$$

注意到 $x_i^T B^{-1} x_i$ 是标量（记为 c），可分离得到：

$$x_i^T A^{-1} = \frac{x_i^T B^{-1}}{1 - x_i^T B^{-1} x_i}$$

- 全样本的正则化系数：$\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T y = B^{-1} X^T y$
- 留一法的正则化系数：$\hat{\beta}_i(\lambda) = (X_{(i)}^T X_{(i)} + \lambda I)^{-1} X_{(i)}^T y_{(i)} = A^{-1} X_{(i)}^T y_{(i)}$

现在计算“留一法预测残差” $y_i - x_i^T \hat{\beta}_i(\lambda)$：

$$
\begin{aligned}
y_i - x_i^T \hat{\beta}_i(\lambda) &= y_i - x_i^T A^{-1} X_{(i)}^T y_{(i)} \\
&= y_i - x_i^T A^{-1} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} x_j y_j \right) \qquad (\text{因 } X_{(i)}^T y_{(i)} = \sum_{\substack{j=1 \\ j \neq i}}^{n} x_j y_j)
\end{aligned}
$$

再计算“全样本预测残差” $y_i - x_i^T \hat{\beta}(\lambda)$:

$$
y_i - x_i^T \hat{\beta}(\lambda) = y_i - x_i^T B^{-1} X^T y = y_i - x_i^T B^{-1} \left(\sum_{j=1}^{n} x_j y_j \right)
$$

结合步骤2中 $x_i^T A^{-1} = \frac{x_i^T B^{-1}}{1 - x_i^T B^{-1} x_i}$，对“留一法预测残差”变形:

$$
\begin{aligned}
y_i - x_i^T \hat{\beta}_i(\lambda) &= y_i - \frac{x_i^T B^{-1}}{1 - x_i^T B^{-1} x_i} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} x_j y_j \right) \\
&= y_i - \frac{x_i^T B^{-1} X^T y - x_i^T B^{-1} x_i y_i}{1 - x_i^T B^{-1} x_i} \qquad (\text{拆分 } \sum_{j=1}^{n} x_j y_j = \sum_{\substack{j=1 \\ j \neq i}}^{n} x_j y_j + x_i y_i) \\
&= \frac{\left(y_i - x_i^T B^{-1} X^T y\right)\left(1 - x_i^T B^{-1} x_i\right) + x_i^T B^{-1} x_i y_i}{1 - x_i^T B^{-1} x_i} \\
&= \frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - x_i^T B^{-1} x_i}
\end{aligned}
$$

注意到 $H_{ii}(\lambda) = x_i^T (X^T X + \lambda I)^{-1} x_i = x_i^T B^{-1} x_i$，因此最终得:

$$
y_i - x_i^T \hat{\beta}_i(\lambda) = \frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - H_{ii}(\lambda)}
$$

广义交叉验证的简化逻辑

由于直接计算每个 $H_{ii}(\lambda)$ 成本较高，利用**迹的线性性**（$\mathrm{tr} H = \sum_{i=1}^{n} H_{ii}(\lambda)$），近似认为每个 $H_{ii}(\lambda) \approx \frac{\mathrm{tr} H}{n}$，从而将损失函数简化为:

$$
\lambda = \arg\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - \frac{\mathrm{tr} H}{n}} \right)^2
$$

AIC

AIC（Akaike Information Criterion）的核心思想是在**极大似然**的基础上，引入**模型复杂度惩罚项**，其定义为：

$$\text{AIC} \triangleq \max_k \left(L_k - k\right)$$

其中，L_k 是模型的**对数似然函数**（log-likelihood），k 是模型中待估参数的数量。

线性回归模型的设定

考虑线性回归模型：

$$y_i = x_i^T \beta + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

写成矩阵形式为：

$$Y = X\beta + \varepsilon$$

其中，$Y = [y_1, y_2, \ldots, y_n]^T$ 是响应向量，$X = [x_1, x_2, \ldots, x_n]^T$ 是设计矩阵（x_i 为第 i 个样本的特征向量），$\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$ 是误差项，且假设 $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$，因此 $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$。

为应对多重共线性，引入**正则化项**（以岭回归为例，罚项为 λ），此时回归系数的估计为：

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

似然函数与对数似然函数

- **似然函数**：基于 $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$，其概率密度函数为：

$$L(\beta, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \|Y - X\beta\|^2\right\}$$

- **对数似然函数**：对似然函数取自然对数，得：

$$\log L(\beta,\sigma) = -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{n}{2}\log\sigma^2 + \text{constant}$$

其中“constant”为与 β 和 σ 无关的常数项。

对σ^2和β的极大似然估计

1. **对σ^2求极大似然估计**:
   将 $\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y$ 代入对数似然函数，对 σ^2 求导并令导数为0，可得:

$$\hat{\sigma}^2 = \frac{1}{n}\|Y - X\hat{\beta}(\lambda)\|^2$$

2. **代入对数似然函数**:
   将 $\hat{\sigma}^2$ 代入对数似然函数，化简后得（忽略常数项）：

$$\max\log L = -\frac{n}{2} - \frac{n}{2}\log\left(\frac{1}{n}\|Y - X\hat{\beta}(\lambda)\|^2\right)$$

在正则化回归中，参数的“有效数量”需通过**帽子矩阵的迹**来定义。记帽子矩阵为:

$$H(\lambda) = X(X^T X + \lambda I)^{-1} X^T$$

则有效参数数量定义为帽子矩阵的迹:

$$k_{\text{eff}} = \text{tr}\left(H(\lambda)\right) = \text{tr}\left(X(X^T X + \lambda I)^{-1} X^T\right)$$

- 当 $\lambda = 0$（无正则化，普通最小二乘）时，$H(0) = X(X^T X)^{-1} X^T$，若 X 列数为 $k+1$（含截距项），则 $\text{tr}(H(0)) = k+1$，与普通线性回归的参数数量一致。
- 当 $\lambda \to \infty$（强正则化）时，$\hat{\beta}(\lambda) \to 0$，此时 $\text{tr}(H(\infty)) \to 0$，符合参数被“压缩至0”的直觉。

正则化下的AIC结论

结合“对数似然极大化”和“模型复杂度惩罚”，正则化线性回归的AIC为:

$$\text{AIC} \triangleq -\frac{n}{2}\log\|Y - X\hat{\beta}(\lambda)\|_2^2 - \text{tr}\left(X(X^T X + \lambda I)^{-1} X^T\right)$$

最终，正则化参数 λ 的选择需最大化AIC，即:

$$\lambda = \arg\max_{\lambda} \mathrm{AIC}$$