ASTA6002
回归分析与数据思维

高维线性回归与正则化

中国科学技术大学
管理学院统计与金融系　曾靖

学习目标

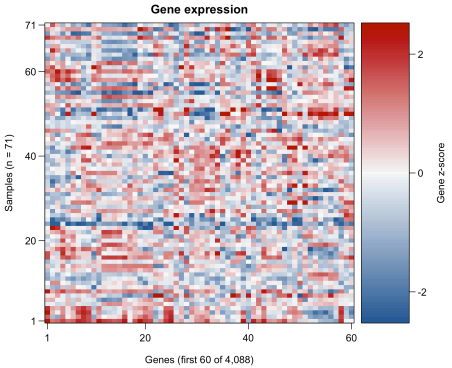
- **认识高维回归问题：**从实际案例出发，解释最小二乘估计的不可识别性与不稳定性。
- **理解正则化与稀疏估计：**掌握 Ridge 与 Lasso 的收缩机制，理解偏差、方差与稀疏性的关系。
- **认识变量选择方法：**理解稀疏模型、活动集与最佳子集选择，了解经典搜索与现代算法。
- **掌握优化思路：**理解软阈值与 KKT 条件，实现坐标下降、近端梯度及正则化路径计算。
- **完成调参与方法比较：**使用五折 CV 及 Min、1-SE 规则选择参数，比较稀疏与密集信号下的估计方法。

真实案例 1：用基因表达预测维生素 B2 生产

- **研究任务：**利用枯草芽孢杆菌的基因表达，预测核黄素（维生素 B2）的生产速率。
- **公开数据：**$n = 71$ 个样本，每个样本有 $p = 4088$ 个基因表达特征，天然满足 $p \gg n$。
- **回归目标：**y 为对数生产速率，X 的各列为基因的对数表达水平。

$$y = X\beta^* + \varepsilon, \qquad X \in \mathbb{R}^{71\times 4088}.$$

思考：如何预测新样本？能否从数千个基因中选出少数有用的预测变量？



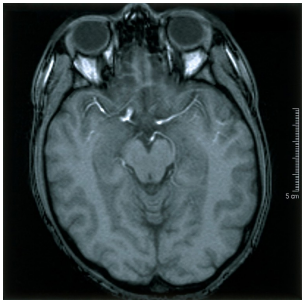
真实数据：展示前 60 个基因

真实案例 2：用脑部 MRI 预测年龄

- **预测任务：**用脑部 MRI 影像预测受试者的年龄 y_i。
- **特征形成：**将影像配准到共同空间，每个体素（voxel）的灰质测量值作为一个特征。
- **维度示例：**$n = 200$ 位受试者，保留 $p = 10^5$ 个体素，满足 $p \gg n$。

$$X \in \mathbb{R}^{200\times 10^5}, \qquad y \in \mathbb{R}^{200}.$$

思考：相邻体素高度相关。为每个体素估计一个系数，结果会稳定吗？



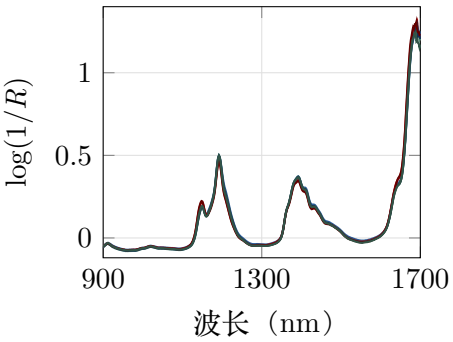
脑部 MRI 切片（影像示例）

三维影像 $\longrightarrow$ $x_i \in \mathbb{R}^p$

真实案例 3：用近红外光谱预测汽油辛烷值

- **预测任务：**用近红外光谱（near-infrared, NIR）预测汽油的辛烷值。
- **公开数据：**$n = 60$ 个汽油样本，测量 900–1700 nm，每隔 2 nm 记录一次，得到 $p = 401$ 个波长特征。
- **特征形成：**$X_{ij} = \log(1/R_{ij})$，R_{ij} 为反射率；相邻波长的测量往往高度相关。

$$X \in \mathbb{R}^{60\times 401}, \qquad p > n.$$



真实数据：选取 6 个样本

思考：选择少数波长，还是利用整条光谱？如何控制冗余信息？

真实案例 4：用文本词频预测影评分数

- **预测任务：**根据一段影评，预测读者给出的数值评分 y_i。
- **特征形成：**词袋模型（bag-of-words）将每个词作为一列，以它在文档中的出现次数作为特征。
- **维度示例：**$n = 500$ 篇带评分影评，词表含 $p = 10^4$ 个词，满足 $p \gg n$。
  $X \in \mathbb{R}^{500\times 10^4}, \quad X_{ij} = \mathrm{count}(w_j, d_i).$

思考：多数词只在少数影评中出现。哪些词的系数能够被可靠地估计？

d_1： 精彩 自然
d_2： 拖沓 平淡
d_3： 精彩 温暖

$\Downarrow$ 统计每个词的出现次数

	精彩	自然	拖沓	平淡	温暖	无聊
d_1	1	1	0	0	0	0
d_2	0	0	1	1	0	0
d_3	1	0	0	0	1	0

文档—词矩阵（示意）

设计矩阵（design matrix）$X \in \mathbb{R}^{n \times p}$，有 n 行、p 列。

- 若 $p > n$，则 $X^\top X$ 必奇异（singular），普通逆不存在。

$$\operatorname{rank}(X) \le n < p.$$

- 若 $v \neq 0$ 且 $Xv = 0$，则 β 与 $\beta + v$ 的训练拟合值相同：

$$X(\beta + v) = X\beta.$$

训练样本无法区分这两个系数向量，但它们在新样本上的预测可能不同。

最小二乘估计与中心化

- 普通最小二乘（ordinary least squares, OLS）最小化平方损失（squared loss）：

$$\min_{\beta_0,\,\beta} \frac{1}{2n}\|y-\beta_0\mathbf{1}_n - X\beta\|_2^2.$$

- 固定 β，对截距求导得 $\widehat{\beta}_0 = \bar{y} - \bar{x}^\top\beta$，于是

$$y_i - \widehat{\beta}_0 - x_i^\top\beta = (y_i-\bar{y}) - (x_i-\bar{x})^\top\beta.$$

- 令 $X_c = X - \mathbf{1}_n\bar{x}^\top$、$y_c = y - \bar{y}\mathbf{1}_n$，即中心化（centering）。若 X_c 列满秩，则

$$\widehat{\beta}_{\mathrm{OLS}} = \arg\min_{\beta} \frac{\|y_c - X_c\beta\|_2^2}{2n} = (X_c^\top X_c)^{-1}X_c^\top y_c.$$

- 后续用 X, y 表示中心化后的数据，不再单列截距。

Remark 也可令 $X_{\mathrm{aug}} = [\mathbf{1}_n,\ X]$、$\theta = (\beta_0, \beta^\top)^\top$，
即将全一向量并入设计矩阵。此时惩罚只作用于 θ 中的斜率部分。

课堂练习 1：高维回归问题

取 $n = 80$、$p = 500$，X_{ij} 独立服从 $N(0,1)$，生成

$$y = X\beta^* + \varepsilon, \quad \beta^* = (2,2,2,2,2,0,\ldots,0)^\top, \quad \varepsilon \sim N(0, I_n).$$

生成训练集，将 X 各列与 y 中心化，再用普通线性回归估计系数。

- 比较直接求逆与最小二乘求解器的结果。检查矩阵的秩及系数的唯一性。
- 构造 $v \neq 0$、$Xv = 0$，比较 $\widehat{\beta}$ 与 $\widehat{\beta} + v$ 的系数及训练拟合，解释为什么系数不同却能同样拟合数据。

小特征值为何放大方差

即使 $p < n$，较小的特征值也会使相应方向的系数估计方差很大。

- 设 $y = X\beta^* + \varepsilon$，$X$ 列满秩，且 $\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$。

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}} \mid X) = \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.$$

- 对 $X^\top X$ 作特征值分解（eigendecomposition）：

$$X^\top X = Q\,\mathrm{diag}(a_1, \ldots, a_p)\, Q^\top, \qquad a_j > 0.$$

其中 q_j 为单位特征向量（eigenvector），因此

$$\mathrm{Var}(q_j^\top \widehat{\beta}_{\mathrm{OLS}} \mid X) = q_j^\top \sigma^2 (X^\top X)^{-1} q_j = \frac{\sigma^2}{a_j}.$$

- 因此，a_j 越小，沿 q_j 方向的估计方差（variance）越大。

正则化（regularization）

最小二乘只追求拟合训练数据。为控制系数的复杂度，在平方损失上加入惩罚：

$$\widehat{\beta}_\lambda = \arg\min_{\beta} \left\{ \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\text{拟合损失}} + \lambda \underbrace{P(\beta)}_{\text{复杂度惩罚}} \right\}.$$

- $P(\beta)$ 为惩罚项（penalty），表达对系数结构的约束。
- $\lambda \geq 0$ 为调参参数（tuning parameter），控制拟合与惩罚的相对强度。
- $\lambda = 0$ 回到最小二乘；不同的 $P(\beta)$ 引出 Ridge、Lasso 等方法。

Ridge 的二次惩罚与闭式解

- Ridge 回归（ridge regression）取二次惩罚 $P(\beta) = \frac{1}{2}\|\beta\|_2^2$，限制系数的整体大小，得到 Ridge 估计：

$$\widehat{\beta}_R = \arg\min_{\beta} \left\{ \frac{\|y - X\beta\|_2^2}{2n} + \frac{\lambda}{2}\|\beta\|_2^2 \right\} = (X^\top X + n\lambda I)^{-1} X^\top y$$

- $\lambda > 0$ 使矩阵 $X^\top X + n\lambda I$ 正定
- 所有方向均收缩，通常不产生精确零系数

令 $r = \operatorname{rank}(X)$。thin SVD（thin singular value decomposition）只保留非零奇异值（singular values）：

$$X = UDV^\top, \quad U \in \mathbb{R}^{n\times r}, \quad D \in \mathbb{R}^{r\times r}, \quad V \in \mathbb{R}^{p\times r},$$

$$U^\top U = V^\top V = I_r, \qquad D = \operatorname{diag}(d_1, \ldots, d_r), \quad d_k > 0.$$

OLS 系数估计：秩亏时取最小欧氏范数解，得到

$$\widehat{\beta}_{\mathrm{OLS}} = VD^{-1}U^\top y = V\operatorname{diag}\left(\frac{1}{d_k}\right)U^\top y.$$

Ridge 系数估计： 取 $\lambda > 0$，由

$$(X^\top X + n\lambda I_p)\, V = V(D^2 + n\lambda I_r)$$

代入闭式解，可得

$$\begin{aligned}\widehat{\beta}_R &= V(D^2 + n\lambda I_r)^{-1} D\, U^\top y \\ &= V \operatorname{diag}\left(\frac{d_k}{d_k^2 + n\lambda}\right) U^\top y.\end{aligned}$$

拟合响应（fitted response）： 分别左乘 $X = UDV^\top$，得到

$$\begin{aligned}\widehat{y}_{\mathrm{OLS}} &= X\widehat{\beta}_{\mathrm{OLS}} = UU^\top y, \\ \widehat{y}_R &= X\widehat{\beta}_R = U \operatorname{diag}\left(\frac{d_k^2}{d_k^2 + n\lambda}\right) U^\top y.\end{aligned}$$

标准化与 Ridge 的坐标收缩

这里的标准化（standardization）设计满足

$$\frac{X^\top X}{n} = I_p, \qquad z = \frac{X^\top y}{n}.$$

代入 OLS 与 Ridge 的闭式解，得到

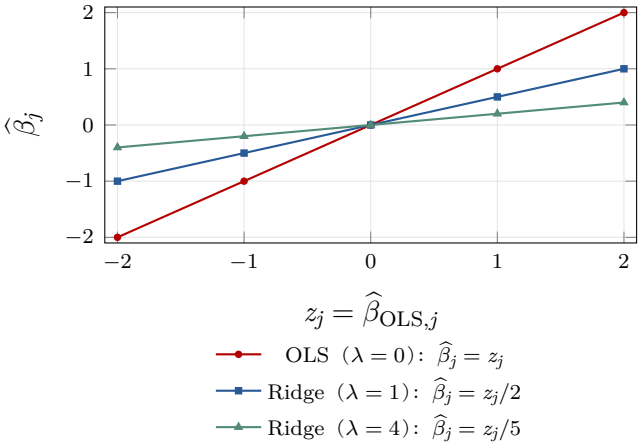
$$\widehat{\beta}_{\mathrm{OLS}} = z, \qquad \widehat{\beta}_R = (I_p + \lambda I_p)^{-1} z = \frac{1}{1+\lambda}\, z.$$

因此，每一个坐标都按相同的比例收缩：

$$\boxed{\widehat{\beta}_{R,j} = \frac{1}{1+\lambda}\widehat{\beta}_{\mathrm{OLS},j}}, \qquad j = 1, \ldots, p.$$

OLS 与 Ridge：系数如何收缩

在 $X^\top X/n = I_p$ 的条件下，令 $z_j = x_j^\top y/n = \widehat{\beta}_{\mathrm{OLS},j}$。



- $\lambda = 0$ 时与 OLS 相同；$\lambda = 1$ 时各系数减半。
- λ 越大，系数越接近零；非零系数不会因有限的 λ 恰好变为零。

$$\widehat{\beta}_{R,j} = \frac{z_j}{1+\lambda}.$$

思考：收缩会引入怎样的偏差

设 $y = X\beta^* + \varepsilon$，且 $E(\varepsilon \mid X) = 0$。

普通最小二乘（ordinary least squares, OLS）：若 X 列满秩，则

$$E(\widehat{\beta}_{\mathrm{OLS}} \mid X) = (X^\top X)^{-1} X^\top X\beta^* = \beta^*, \qquad \mathrm{Bias}(\widehat{\beta}_{\mathrm{OLS}} \mid X) = 0.$$

Ridge：取 $\lambda > 0$，则

$$E(\widehat{\beta}_R \mid X) = (X^\top X + n\lambda I_p)^{-1} X^\top X\beta^*,$$

$$\mathrm{Bias}(\widehat{\beta}_R \mid X) = E(\widehat{\beta}_R \mid X) - \beta^* = -n\lambda(X^\top X + n\lambda I_p)^{-1}\beta^*.$$

- Ridge 一般有偏，但可能降低方差
- 判断优劣应看目标风险，而不能只比较无偏性

课堂练习 2：比较系数收缩

取 $n = 100$、$p = 10$，构建中心化的标准化设计 $X^\top X/n = I_p$，并生成

$$y = X\beta^* + \varepsilon, \qquad \beta^* = (3, 2, 1, 0, \ldots, 0)^\top, \qquad \varepsilon \sim N(0, I_n).$$

在同一数据上拟合 OLS 与不同 λ 的 Ridge。

- 绘制系数路径，将 Ridge 与 OLS 系数并排展示。验证 $\widehat{\beta}_{R,j} = \widehat{\beta}_{\mathrm{OLS},j}/(1+\lambda)$。
- 改用相关设计，仅将各列缩放至单位方差。上式是否成立？每个系数绝对值是否都随 λ 增大而减小？

Lasso：生活中的变量选择——房价预测

- **预测目标：**一套房屋的成交价格
- **候选信息：**面积、区位、楼龄、楼层、朝向、装修，以及周边交通与配套等
- **变量选择（variable selection）：**大量指标可能重复，收集与维护也有成本。能否用少数指标保留主要预测信息？

思考：如果只能保留五项信息，你会选什么？怎样用数据检验你的选择？



面积　区位　楼龄
楼层　朝向　装修

↓

少数指标 ⟶ 房价

生活中的变量选择——预测睡眠质量

- **预测目标：**第二天的睡眠质量评分
- **候选信息：**活动量、心率、入睡时间、屏幕使用、咖啡摄入及大量分时统计量
- **稀疏建模的设想：**在给定候选指标中，少数指标可能已概括主要预测信息，其余提供的增量有限

思考：全天步数与各时段步数高度相关时，应该保留哪些指标？选择结果会稳定吗？



活动　心率　作息

↓

少数指标 ⟶ 评分

稀疏线性模型与活动集

- 设响应由线性模型描述：

$$y = X\beta^* + \varepsilon, \qquad X \in \mathbb{R}^{n\times p}, \quad \beta^* \in \mathbb{R}^p.$$

- **稀疏性（sparsity）：**若 β^* 只有少数非零坐标，则称其为稀疏系数向量。
- 定义真实活动集（active set）与稀疏度（sparsity level）为

$$S = \{j : \beta_j^* \neq 0\}, \qquad s = |S| \ll p.$$

稀疏模型中的变量选择

只有活动集 S 中的变量进入模型：

$$y = \sum_{j \in S} x_j \beta_j^* + \varepsilon = X_S \beta_S^* + \varepsilon,$$

$$X_S \in \mathbb{R}^{n \times s}, \qquad \beta_S^* \in \mathbb{R}^s.$$

- 变量选择的目标：根据数据估计未知的 S。
- 稀疏性是一项建模假设，依赖候选变量的表示。
- Ridge 通常保留所有变量；Lasso 可将部分系数设为零。

最佳子集选择：直接限制变量个数

定义 $\|\beta\|_0 = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}$，即非零系数的个数。最佳子集选择（best subset selection）在给定大小 k 下求解

$$\widehat{\beta}^{(k)} \in \arg\min_{\beta} \frac{\|y - X\beta\|_2^2}{2n} \quad \text{s.t.} \quad \|\beta\|_0 \leq k.$$

对候选子集 $A \subseteq \{1, \ldots, p\}$，记

$$J(A) = \min_{b \in \mathbb{R}^{|A|}} \frac{\|y - X_A b\|_2^2}{2n}, \qquad \widehat{A}_k \in \arg\min_{|A| \leq k} J(A).$$

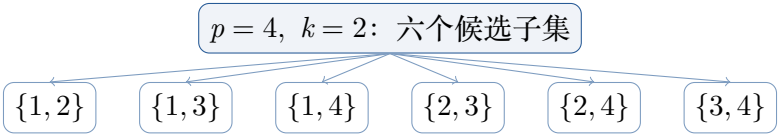
- 也可最小化 $\|y - X\beta\|_2^2/(2n) + \lambda\|\beta\|_0$。
- k 或 λ 仍需选择；训练拟合更好，不保证预测更好。

经典精确搜索：穷举

- **穷举（exhaustive search）：**给定子集大小 k，逐个计算所有候选子集的 $J(A)$，选取最小者。
- **候选数与上界：**对 $1 \le k \le p$，

$$N_k = \binom{p}{k} \le \frac{p^k}{k!} \le \left(\frac{ep}{k}\right)^k.$$

固定 k、p 较大时，$\binom{p}{k} \approx p^k/k!$。



例如 $p = 100$、$k = 5$ 时，$\binom{100}{5} = 75{,}287{,}520 \approx 7.53 \times 10^7$。

经典近似搜索：向前加入与向后删除

向前逐步选择（forward stepwise selection）： 从 $A_0 = \varnothing$ 出发，每次加入使拟合损失最小的一个变量：

$$j_t \in \arg\min_{j \notin A_t} J(A_t \cup \{j\}), \qquad A_{t+1} = A_t \cup \{j_t\}.$$

停止： 达到目标大小 $|A_t| = k$ 时停止加入。

向后逐步选择（backward stepwise selection）： 从可拟合的完整模型出发，每次删除损失增加最小的一个变量：

$$j_t \in \arg\min_{j \in A_t} J(A_t \setminus \{j\}), \qquad A_{t+1} = A_t \setminus \{j_t\}.$$

停止： 达到目标大小 $|A_t| = k$ 时停止删除。

若 k 未知，用 CV 从候选路径中选择模型大小 $\widehat{k}$。

现代精确优化：用整数变量表示选择

Bertsimas、King、Mazumder（2016）用混合整数优化（mixed-integer optimization, MIO）求最佳子集。令 $u_j \in \{0,1\}$ 表示是否允许变量 j 入选，$M > 0$ 为足够大的有效系数界：

$$
\begin{aligned}
\min_{\beta,\, u} \quad & \frac{\|y - X\beta\|_2^2}{2n} \\
\text{s.t.} \quad & -Mu_j \le \beta_j \le Mu_j, \quad j = 1, \ldots, p, \\
& \sum_{j=1}^{p} u_j \le k, \qquad u \in \{0,1\}^p.
\end{aligned}
$$

快速算法：迭代硬阈值（IHT）

迭代硬阈值（iterative hard thresholding, IHT）：先走平方损失的梯度步，再保留绝对值最大的 k 个坐标：

$$w^{(t)} = \beta^{(t)} + \eta X^\top (y - X\beta^{(t)})/n, \qquad \beta^{(t+1)} = H_k\big(w^{(t)}\big)\,.$$

$H_k(w)$ 将其余坐标置零，是到 $\{\beta : \|\beta\|_0 \le k\}$ 的欧氏投影。

ABESS：用 splicing 交换变量

Zhu、Wen 等（2020，PNAS）固定 $|A| = k$，反复交换集合内外的变量。以下令 $x_j^\top x_j / n = 1$。

输入与初始化： X, y, k、交换上限与阈值 $\tau > 0$；给定 $|A| = k$。

1. 在 A 上拟合 OLS，得到 $\widehat{\beta}$、$r = y - X\widehat{\beta}$。
2. 集合内按 $\widehat{\beta}_j^2$ 从小到大、集合外按 $(x_j^\top r/n)^2$ 从大到小排序，构造等量的 A_{out} 与 A_{in}。
3. 尝试不同交换大小，重新拟合候选集合
$$A' = (A \setminus A_{\text{out}}) \cup A_{\text{in}}.$$
4. 若 $J(A) - J(A') > \tau$，接受改进并继续；无可接受交换时停止，输出集合与系数。

多项式时间：哪些条件下成立?

- **一般输入：**最佳子集选择是 NP-hard。
- **高维统计模型：**ABESS 在原文假设下具有多项式复杂度，并有高概率最优子集与真实支持恢复保证。

其中设计条件涉及稀疏方向上的谱界及交叉相关控制，例如

$$c_-\|v\|_2^2 \le \frac{\|Xv\|_2^2}{n} \le c_+\|v\|_2^2, \quad \|v\|_0 \le 2k, \quad 0 < c_- \le c_+.$$

研究拓展：更精细的交换与更一般的损失

Optimal Pursuit：Zhu、Zhang、Xia（ICML 2025）

- 对变量进入和退出时的子问题作更精确的优化，更新选择与删除准则。
- 改善的是当前进入、退出决策；“当前步骤最优”不等于对任意输入的全局最优。

SCOPE：Zhu 等（2026 修订预印本，arXiv v2）

- 用目标函数引导 splicing，将稀疏约束优化拓展到一般可微损失。
- 在低维子空间的强凸性、光滑性等条件下，研究线性收敛和正确稀疏度下的支持恢复。

比较算法时，区分计算复杂度、优化误差、支持恢复与预测误差。

课堂练习 3：比较子集搜索

固定种子，取 $n = 80, p = 16, k = 3$，*生成训练数据并中心化：*

$$X_i \sim N(0, \Sigma), \quad \Sigma_{jl} = 0.2^{|j-l|}, \quad y = X\beta^* + \varepsilon,$$

$\beta^* = (2, 2, 2, 0, \ldots, 0)^\top$，$\varepsilon \sim N(0, I_n)$。

- 穷举 $\binom{16}{3} = 560$ 个子集作为基准；比较向前选择、IHT 与 ABESS 的 $J(A)$、入选集合和运行时间。

从最佳子集选择到 Lasso

ℓ_0 惩罚对每个非零系数收取固定代价；Lasso 用绝对值总量替代计数：

$$\widehat{\beta}_{\ell_0} \in \arg\min_{\beta} \left\{ \frac{\|y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_0 \right\},$$

$$\widehat{\beta}_{L} \in \arg\min_{\beta} \left\{ \frac{\|y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_1 \right\}.$$

- 最佳子集直接控制非零个数，涉及组合搜索。
- Lasso 的目标是凸函数，同时产生连续收缩与稀疏解。
- 两者的优化难度、估计偏差和变量选择行为不同。

Lasso 的名称与由来

Lasso 是 *Least Absolute Shrinkage and Selection Operator* 的缩写。
由 Robert Tibshirani（1996）提出。



Robert Tibshirani



lasso：套索
绳圈可收紧，用于套住马、牛等

- **Absolute：** 限制系数的绝对值总量。
- **Shrinkage and Selection：** 同时实现系数收缩与变量选择。

Lasso 估计量：约束形式与惩罚形式

约束形式（constrained form）：限制系数绝对值之和。

$$\widehat{\beta}_L(t) \in \arg\min_{\beta} \frac{\|y - X\beta\|_2^2}{2n} \quad \text{s.t.} \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \le t.$$

$t \ge 0$ 控制系数绝对值总量。

惩罚形式（penalized form）：在平方损失中加入 ℓ_1 惩罚。

$$\widehat{\beta}_L(\lambda) \in \arg\min_{\beta} \left\{ \frac{\|y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_1 \right\}.$$

$\lambda \ge 0$ 控制惩罚强度。

标准化设计下：OLS、Ridge 与 Lasso

设 X, y 已中心化，且 X 满足标准化条件

$$X^\top X/n = I_p, \qquad z = X^\top y/n.$$

Lasso 目标展开并略去常数后，按坐标分离：

$$\min_\beta \sum_{j=1}^{p} \left\{ \frac{1}{2}(\beta_j - z_j)^2 + \lambda|\beta_j| \right\}.$$

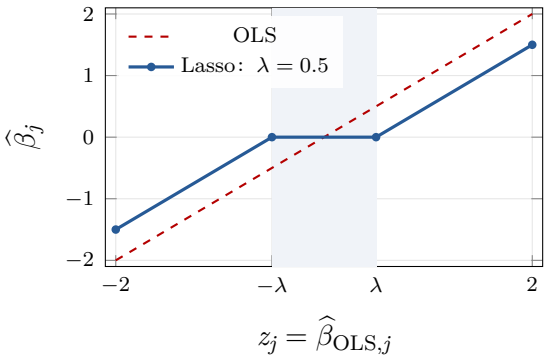
三种估计的坐标形式为

$$\widehat{\beta}_{\mathrm{OLS},j} = z_j, \qquad \widehat{\beta}_{R,j} = \frac{z_j}{1+\lambda}, \qquad \widehat{\beta}_{L,j} = S_\lambda(z_j).$$

- Ridge 按比例缩小每个系数。
- Lasso 的软阈值（soft-thresholding）使小系数归零。

软阈值函数 $S_\lambda(z)$

在标准化正交设计下，$z_j = \widehat{\beta}_{\mathrm{OLS},j}$，Lasso 系数为 $S_\lambda(z_j)$。



$$S_\lambda(z) = \begin{cases} z - \lambda, & z > \lambda, \\ 0, & |z| \le \lambda, \\ z + \lambda, & z < -\lambda. \end{cases}$$

- 当 $\lambda > 0$ 时，所有非零 OLS 系数都向零收缩。
- $|\widehat{\beta}_{\mathrm{OLS},j}| \le \lambda$ 时，系数直接被压到零。
- 超过阈值时，系数绝对值减少 λ，符号保持不变。

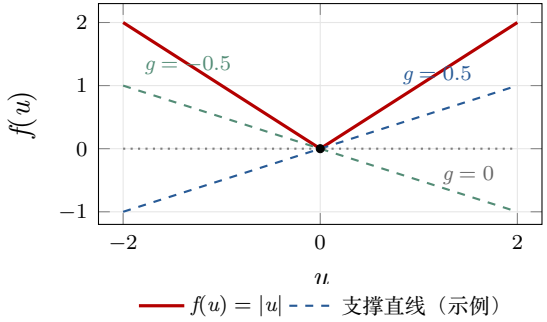
次梯度：用支撑直线描述不可导点

设 $f\colon \mathbb{R} \to \mathbb{R}$ 为凸函数。若 g 满足

$$f(u) \geq f(b) + g(u - b), \qquad \forall u \in \mathbb{R},$$

称 g 为 b 处的**次梯度（subgradient）**。

所有次梯度组成**次微分（subdifferential）** $\partial f(b)$。



连续凸函数 $f(u) = |u|$ 的次微分为

$$\partial f(b) = \begin{cases} \{1\}, & b > 0, \\ [-1, 1], & b = 0, \\ \{-1\}, & b < 0. \end{cases}$$

图中虚线均过原点，且始终不高于实线；零点的支撑斜率可以不唯一。

用次梯度求解一维 Lasso

令 $f(b)=\frac{1}{2}(b-z)^2+\lambda|b|$，$\lambda>0$。由凸性，最优性的充要条件是

$$0\in\partial f(\widehat{b})=\widehat{b}-z+\lambda\partial|\widehat{b}|.$$

1. **若** $\widehat{b}>0$：$\partial|\widehat{b}|=\{1\}$，于是

$$0=\widehat{b}-z+\lambda\implies\widehat{b}=z-\lambda>0\iff z>\lambda.$$

2. **若** $\widehat{b}<0$：$\partial|\widehat{b}|=\{-1\}$，于是

$$0=\widehat{b}-z-\lambda\implies\widehat{b}=z+\lambda<0\iff z<-\lambda.$$

3. **若** $\widehat{b}=0$：$\partial|\widehat{b}|=[-1,1]$，于是

$$0\in-z+\lambda[-1,1]\iff z\in[-\lambda,\lambda]\iff|z|\le\lambda.$$

因此 $\widehat{b}=S_\lambda(z)$；正交设计下，$\widehat{\beta}_{L,j}=S_\lambda(z_j)$。

一般设计下的 KKT 条件

KKT（Karush–Kuhn–Tucker）条件适用于一般设计。Lasso 目标为

$$J(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad \lambda > 0.$$

第一步：记平方损失为 $L(\beta)$，对其求梯度：

$$\nabla L(\beta) = -\frac{1}{n} X^\top (y - X\beta), \qquad \frac{\partial L}{\partial \beta_j} = -\frac{x_j^\top (y - X\beta)}{n}.$$

第二步：由凸性，最优解满足 $0 \in \partial J(\widehat{\beta})$。存在 $g_j \in \partial |\widehat{\beta}_j|$，使

$$0 = -\frac{x_j^\top (y - X\widehat{\beta})}{n} + \lambda g_j.$$

记 $\widehat{r} = y - X\widehat{\beta}$，整理得到 $x_j^\top \widehat{r}/n = \lambda g_j$。

代入绝对值的次梯度，将最优性条件分成两种情况：

1. **若** $\widehat{\beta}_j \neq 0$，则 $g_j = \operatorname{sign}(\widehat{\beta}_j)$，所以

$$\frac{x_j^\top (y - X\widehat{\beta})}{n} = \lambda \operatorname{sign}(\widehat{\beta}_j).$$

正系数对应 $+\lambda$，负系数对应 $-\lambda$。

2. **若** $\widehat{\beta}_j = 0$，则 $g_j \in [-1, 1]$，所以

$$\frac{x_j^\top (y - X\widehat{\beta})}{n} \in [-\lambda, \lambda] \quad \Longleftrightarrow \quad \left| \frac{x_j^\top (y - X\widehat{\beta})}{n} \right| \leq \lambda.$$

残差与该变量的内积绝对值不超过阈值，系数可以保持为零。

算法检查：若某个零系数满足 $|x_j^\top (y - X\widehat{\beta})/n| > \lambda$，当前解就不是最优解。

- 对中心化数据，将 $\beta = 0$ 代入 KKT

$$\lambda_{\max} = \frac{\|X^\top y\|_\infty}{n}$$

- λ 足够大时所有斜率为零
- 正则化路径通常从此处逐步减小 λ

课堂练习 4：验证软阈值与 KKT

取 $n = 100$、$p = 10$，构建中心化的标准化设计 $X^\top X/n = I_p$，生成

$$y = X\beta^* + \varepsilon, \quad \beta^* = (2, -1, 0.5, 0, \ldots, 0)^\top, \quad \varepsilon \sim N(0, I_n).$$

在同一数据上计算 $\lambda = 0.3$ 时的 *Lasso* 与 *Ridge* 估计。

- 令 $z = X^\top y/n$，比较 Lasso 估计 $\widehat{\beta}_L = S_\lambda(z)$ 与 Ridge 估计 $\widehat{\beta}_R = z/(1+\lambda)$，画系数对比图并报告各自的估计活动集。
  回顾： 估计活动集 $\widehat{S} = \{j : \widehat{\beta}_j \neq 0\}$；真实活动集 $S = \{j : \beta^*_j \neq 0\}$。
- 计算 Lasso 残差 $\widehat{r}$ 及 $c_j = x_j^\top \widehat{r}/n$，逐坐标检查非零系数的符号等式和零系数的不等式。

一般设计矩阵下如何计算 Lasso

对中心化数据求解 Lasso 目标：

$$\min_{\beta} J(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

优化思路：每次固定其他坐标，仅优化一个系数 β_j：

$$\beta_j^{\text{new}} \in \arg\min_{b \in \mathbb{R}} J(\beta_1, \ldots, \beta_{j-1}, b, \beta_{j+1}, \ldots, \beta_p).$$

这样将一个多维优化问题，转化为一系列一维优化问题。

循环坐标下降：优化思路

循环坐标下降（coordinate descent）

给定初始系数，重复以下过程：

1. 固定其余系数，只把 β_j 当作未知量，精确求解一维问题。
2. 立即使用更新后的系数，按 $j = 1, \ldots, p$ 依次更新。
3. 一轮结束后检查 KKT；若未满足，则继续下一轮。

每次单坐标最小化都使目标值 $J(\beta)$ 不增。

固定当前其余系数，令 b 替代第 j 个系数。从响应中减去其他变量的贡献，定义**部分残差（partial residual）**：

$$r^{(j)} = y - \sum_{k \neq j} x_k \beta_k.$$

仅与 b 有关的目标为

$$\min_b \frac{1}{2n} \| r^{(j)} - x_j b \|_2^2 + \lambda |b|.$$

若已保存完整残差 $r = y - X\beta$，则

$$r^{(j)} = r + x_j \beta_j.$$

其他坐标使用当前最新值，因此每一步都接续前一步的结果。

令 $a_j = x_j^\top x_j/n > 0$，$c_j = x_j^\top r^{(j)}/n$。展开单坐标目标并略去与 b 无关的常数：

$$\min_b \left\{ \frac{a_j}{2} b^2 - c_j b + \lambda |b| \right\} = \min_b \left\{ \frac{a_j}{2} \left(b - \frac{c_j}{a_j} \right)^2 + \lambda |b| \right\}.$$

除以 a_j 后应用一维软阈值公式：

$$\beta_j^{\text{new}} = S_{\lambda/a_j}(c_j/a_j) = \frac{S_\lambda(c_j)}{a_j}.$$

- 若各列缩放至 $x_j^\top x_j/n = 1$，更新简化为 $S_\lambda(c_j)$
- 此处不要求列之间正交；相关性通过部分残差进入 c_j

算法：Lasso 的循环坐标下降

输入： $X, y, \lambda > 0$、容差 $\epsilon > 0$、最大轮数 T；预先计算 $a_j = \|x_j\|_2^2/n > 0$。

1. 初始化： $\beta = 0$，$r = y$。

2. 循环更新： 每轮按 $j = 1, \dots, p$ 顺序执行

$$b_{\text{old}} \leftarrow \beta_j, \qquad c_j \leftarrow x_j^\top (r + x_j b_{\text{old}})/n,$$
$$\beta_j \leftarrow S_\lambda(c_j)/a_j, \qquad r \leftarrow r + x_j(b_{\text{old}} - \beta_j).$$

3. 检查停止： 一轮后令 $q = X^\top r/n$，检查

$$q_j = \lambda \operatorname{sign}(\beta_j) \ (\beta_j \neq 0), \qquad |q_j| \le \lambda \ (\beta_j = 0).$$

全部条件在容差 ϵ 内满足则停止，否则继续，最多 T 轮。

输出： 系数、目标值和 KKT 偏离量；达到轮数上限时报告未收敛。

课堂练习 5：相关设计的一轮更新

考虑二维 Lasso 目标 $J(\beta)=\frac{1}{2}\beta^\top G\beta - z^\top\beta + \lambda\|\beta\|_1$，其中

$$G=\frac{X^\top X}{n}=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix},\qquad z=\frac{X^\top y}{n}=\begin{pmatrix}1\\ 0.8\end{pmatrix},\quad \lambda=0.2.$$

从 $\beta^{(0)}=(0,0)^\top$ 开始，按第 1、第 2 坐标的顺序完成一轮循环坐标下降。

- 写出 $c_1=z_1-G_{12}\beta_2^{(0)}$，计算更新后的 $\beta_1^{(1)}$。
- 使用刚得到的 $\beta_1^{(1)}$，计算 $c_2=z_2-G_{21}\beta_1^{(1)}$ 及 $\beta_2^{(1)}$。
- 计算 $q=z-G\beta^{(1)}$，检查一轮更新后的系数是否已满足 KKT。

课堂练习 5 答案：一轮坐标更新与 KKT 检查

1. 更新第一个坐标：

$$c_1 = 1 - 0.5 \times 0 = 1, \qquad \beta_1^{(1)} = S_{0.2}(1) = 0.8.$$

2. 使用最新的 $\beta_1^{(1)}$ 更新第二个坐标：

$$c_2 = 0.8 - 0.5 \times 0.8 = 0.4, \qquad \beta_2^{(1)} = S_{0.2}(0.4) = 0.2.$$

3. 检查 KKT： 一轮后 $\beta^{(1)} = (0.8, 0.2)^\top$，因此

$$q = z - G\beta^{(1)} = \begin{pmatrix} 1 \\ 0.8 \end{pmatrix} - \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}.$$

两个系数均为正，KKT 要求 $q_1 = q_2 = \lambda = 0.2$。
第一个坐标不满足条件，**一轮更新后尚未达到最优解**，需要继续迭代。

近端梯度法：二次近似与梯度步

近端梯度法（proximal gradient method）考虑复合凸优化问题

$$\min_{\beta} J(\beta) = f(\beta) + g(\beta).$$

设 f 可微、凸且梯度为 L-Lipschitz，g 为适当闭凸函数。

先近似光滑部分 f： 在当前点 β^t 构造局部二次模型

$$Q_\eta(u \mid \beta^t) = f(\beta^t) + \langle \nabla f(\beta^t), u - \beta^t \rangle + \frac{1}{2\eta}\|u - \beta^t\|_2^2.$$

对 u 求导并令其为零，得到

$$\nabla f(\beta^t) + \frac{1}{\eta}(u - \beta^t) = 0, \qquad v^t = \arg\min_u Q_\eta(u \mid \beta^t) = \beta^t - \eta \nabla f(\beta^t).$$

二次模型的最优点 v^t，恰好是一步梯度下降（gradient descent）的结果。当 $0 < \eta \le 1/L$ 时，$Q_\eta(u \mid \beta^t) \ge f(u)$。

近端映射与近端梯度更新

近端映射（proximal mapping）定义为

$$\mathrm{prox}_g(v) = \arg\min_u \left\{ g(u) + \frac{1}{2}\|u - v\|_2^2 \right\}.$$

将 g 保留原样：与 f 的二次模型一起最小化，令

$$\beta^{t+1} = \arg\min_u \{ Q_\eta(u \mid \beta^t) + g(u) \}.$$

记 $v^t = \beta^t - \eta \nabla f(\beta^t)$。配方后，存在与 u 无关的 C_t，使

$$Q_\eta(u \mid \beta^t) + g(u) = g(u) + \frac{1}{2\eta}\|u - v^t\|_2^2 + C_t.$$

略去常数，并将目标乘以 η，得到

$$\beta^{t+1} = \arg\min_u \left\{ \eta g(u) + \frac{1}{2}\|u - v^t\|_2^2 \right\} = \mathrm{prox}_{\eta g}(v^t).$$

$$\boxed{\beta^{t+1} = \mathrm{prox}_{\eta g}\big(\beta^t - \eta \nabla f(\beta^t)\big).}$$

记 $\widehat{\Sigma} = X^\top X/n$，将 Lasso 写成 $f(\beta) + g(\beta)$，其中

$$f(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2, \qquad g(\beta) = \lambda\|\beta\|_1.$$

第一步：梯度更新

$$\nabla f(\beta) = -X^\top(y - X\beta)/n, \qquad v^t = \beta^t + \frac{\eta}{n}X^\top(y - X\beta^t).$$

第二步：求近端子问题，各坐标独立：

$$\beta_j^{t+1} = \arg\min_b \left\{\frac{1}{2}(b - v_j^t)^2 + \eta\lambda|b|\right\} = S_{\eta\lambda}(v_j^t).$$

$$\boxed{\beta^{t+1} = S_{\eta\lambda}\Big(\beta^t + \frac{\eta}{n}X^\top(y - X\beta^t)\Big)}, \qquad 0 < \eta \le \frac{1}{\|\widehat{\Sigma}\|_{\mathrm{op}}}.$$

与循环坐标下降相比，这里每轮用同一旧值 β^t 同时更新全部坐标。

正则化路径与热启动

正则化路径（regularization path）沿递减网格求解。记 $\mathcal{A}_\lambda(b)$ 为以 b 为初值的求解器，从全零初值起步：

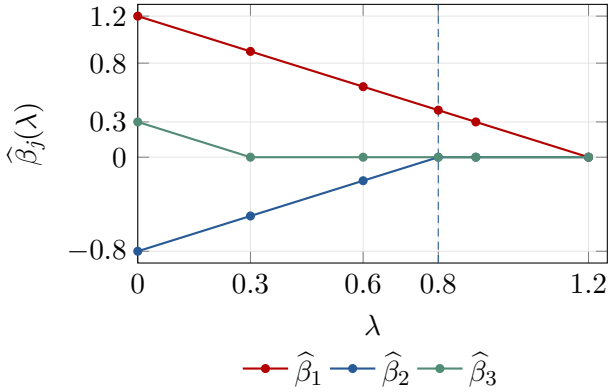
$$\lambda_1 = \lambda_{\max} = \|X^\top y\|_\infty / n, \qquad \lambda_1 > \lambda_2 > \cdots > \lambda_M > 0.$$

热启动（warm start）：把上一个问题的解传给下一个问题。

$$\beta_{\text{init}}^{(1)} = 0, \qquad \beta_{\text{init}}^{(m)} = \widehat{\beta}(\lambda_{m-1}), \qquad m = 2, \ldots, M.$$

$$\widehat{\beta}(\lambda_m) = \mathcal{A}_{\lambda_m}\left(\beta_{\text{init}}^{(m)}\right), \qquad m = 1, \ldots, M.$$

- 相邻 λ 的解往往接近，热启动可能减少迭代次数
- **冷启动（cold start）：**每个 λ_m 都从 $\beta = 0$ 开始
- 初值只提供起点；必须继续求解，不能直接当成新问题的解



- 取 $z = (1.2, -0.8, 0.3)^\top$
- 从 $\lambda = 1.2$ 逐步降到 0 时，不同变量依次进入
- 变点依次为 1.2、0.8、0.3，由 $|z_j|$ 决定

$$\widehat{\beta}_j(\lambda) = \text{sign}(z_j)(|z_j| - \lambda)_+$$

课堂练习 6：两个系数的正则化路径

设 $X^\top X/n = I_2$，$z = X^\top y/n = (1, -0.5)^\top$。分别求解 $\lambda = 1$ 与 $\lambda = 0.5$ 时的 Lasso 估计，使用

$$\widehat{\beta}_j(\lambda) = S_\lambda(z_j).$$

- 计算 $\widehat{\beta}(1)$ 与 $\widehat{\beta}(0.5)$。
- 在 $0 \le \lambda \le 1.2$ 上，画出 $\widehat{\beta}_1(\lambda)$ 与 $\widehat{\beta}_2(\lambda)$ 的两条折线，并标出各自的变点。

课堂练习 6 答案：两个系数的正则化路径

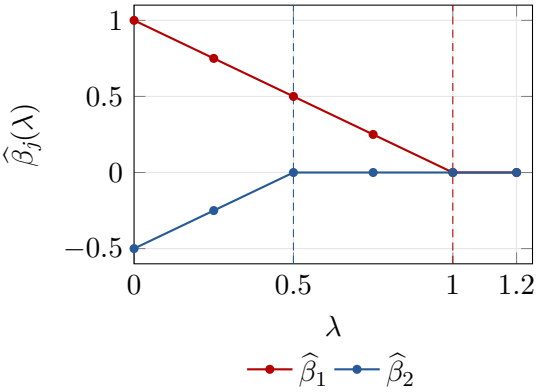
由软阈值公式，两个系数分别为

$$\widehat{\beta}_1(\lambda) = (1-\lambda)_+,$$
$$\widehat{\beta}_2(\lambda) = -(0.5-\lambda)_+.$$

因此，两组估计为

$$\widehat{\beta}(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \widehat{\beta}(0.5) = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}.$$

两个系数的变点分别为 $\lambda = 1$ 和 $\lambda = 0.5$。



课堂练习 7：冷启动与热启动

生成 $n = 120$、$p = 200$ 的相关高维数据，设 X 的行独立服从 $N(0, \Sigma)$，$\Sigma_{jk} = 0.7^{|j-k|}$。中心化并逐列标准化，使 $X_j^\top X_j / n = 1$，再令

$$y = X\beta^* + \varepsilon, \quad \beta_j^* = 2 \ (j \le 5), \quad \beta_j^* = 0 \ (j > 5), \quad \varepsilon \sim N(0, I_n).$$

实现循环坐标下降，比较同一递减网格上的两种启动方式。

- 取 30 个对数等距点 $\lambda_m = \lambda_{\max}(0.05)^{(m-1)/29}$（$m = 1, \dots, 30$），即相邻点的比值相同。分别冷启动、热启动，每点最多 5000 轮；记录迭代轮数、耗时和目标值，画系数路径及轮数对比图。
- 检查两种方式的拟合与目标值是否一致，报告未收敛点。改变变量相关强度后，加速效果是否变化？

调参的目标：降低泛化误差

记 D 为 m 个训练样本，$\widehat{f}_{\lambda,D}$ 为完整建模流程得到的预测函数。对独立、同分布的新样本 (X_0, Y_0)，定义**条件泛化误差（conditional generalization error）**

$$R(\lambda \mid D) = E\Big[(Y_0 - \widehat{f}_{\lambda,D}(X_0))^2 \mid D\Big].$$

固定训练样本量 m，对随机训练集 D 的分布取期望，得到**平均预测风险（expected prediction risk）**

$$R(\lambda) = E_D[R(\lambda \mid D)].$$

- 调参希望找到样本外预测风险较小的 λ
- 训练误差重复使用拟合数据，通常偏乐观
- 验证方法用未参与当前拟合的样本，近似评价预测风险

五折交叉验证：让每一折轮流做验证

交叉验证（cross-validation, CV）将数据均分为五折（fold）。每次留一折验证、用其余四折训练，记训练集为 $D^{(-k)}$。

	V_1	V_2	V_3	V_4	V_5
第 1 次	验证	训练	训练	训练	训练
第 2 次	训练	验证	训练	训练	训练
第 3 次	训练	训练	验证	训练	训练
第 4 次	训练	训练	训练	验证	训练
第 5 次	训练	训练	训练	训练	验证

$$\widehat{R}(\lambda \mid D^{(-k)}) = \frac{1}{|V_k|} \sum_{i \in V_k} (y_i - \widehat{f}_{\lambda, D^{(-k)}}(x_i))^2.$$

$$\widehat{R}(\lambda) = \frac{1}{5} \sum_{k=1}^{5} \widehat{R}(\lambda \mid D^{(-k)}).$$

怎样计算整条 CV path?

预先给定候选网格 $\Lambda = \{\lambda_1 > \cdots > \lambda_M\}$，固定折划分。

1. 对每一折 k：仅用 $D^{(-k)} = D \setminus V_k$ 计算均值、尺度及其他预处理
2. 在该训练折上沿 $\lambda_1, \ldots, \lambda_M$ 拟合路径，可在**同一折内**热启动
3. 将训练折的预处理用于 V_k，计算每个 $\widehat{R}(\lambda_j \mid D^{(-k)})$
4. 按列汇总折误差，得到 $\widehat{R}(\lambda_j)$；选择后再用全部开发数据重拟合

两个不同的 path： 系数路径是 $\lambda \mapsto \widehat{\beta}(\lambda)$；CV path 是 $\lambda \mapsto \widehat{R}(\lambda)$。

例子

取 $n = 100$，分为五折，每折 20 个样本；每次用 $m = 80$ 个样本拟合。下表是用于演示计算的折内验证 MSE：

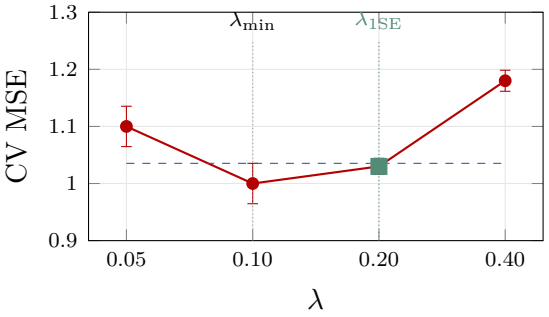
λ	0.05	0.10	0.20	0.40
$\widehat{R}(\lambda \mid D^{(-1)})$	1.20	0.90	0.99	1.13
$\widehat{R}(\lambda \mid D^{(-2)})$	1.10	1.00	1.03	1.18
$\widehat{R}(\lambda \mid D^{(-3)})$	1.00	1.10	1.07	1.23
$\widehat{R}(\lambda \mid D^{(-4)})$	1.15	0.95	1.01	1.15
$\widehat{R}(\lambda \mid D^{(-5)})$	1.05	1.05	1.05	1.21
$\widehat{R}(\lambda)$	1.10	1.00	1.03	1.18

例如，$\widehat{R}(0.10) = (0.90 + 1.00 + 1.10 + 0.95 + 1.05)/5 = 1.00$。

$$\widehat{\lambda}_{\min} \in \arg\min_{\lambda \in \Lambda} \widehat{R}(\lambda) \quad \Longrightarrow \quad \widehat{\lambda}_{\min} = 0.10.$$

这里的“最优”指**候选网格上 CV 误差最小**，并非已知真实风险的最小点。

实例：从 CV error plot 选择 λ



1. **Min 规则：**选择 CV 均值最小的 λ。本例 $\lambda_{\min} = 0.10$。
2. **1-SE 规则：**在最低误差加 1 个标准误的范围内，选择最大的 λ。

$$\widehat{\mathrm{SE}}(\lambda_{\min}) = \sqrt{0.025/20} \approx 0.035.$$

允许上限为 1.035（蓝色虚线），因此 $\lambda_{1\mathrm{SE}} = 0.20$。

$$\lambda_{1\mathrm{SE}} = \max\{\lambda \in \Lambda : \widehat{R}(\lambda) \le \widehat{R}(\lambda_{\min}) + \widehat{\mathrm{SE}}(\lambda_{\min})\}.$$

选定参数后：用全部数据重拟合

由 Min 或 1-SE 规则确定 $\widehat{\lambda}$ 后，固定该参数，拟合最终模型。

交叉验证：选择 $\widehat{\lambda}$ → 全部数据：重新预处理并拟合

$$\widehat{\beta}_{\text{final}} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \widehat{\lambda} P(\beta) \right\}.$$

- 用全部数据重新计算预处理参数，完成中心化与逐列标准化。
- 此处 X, y 表示全部数据经预处理后的设计矩阵和响应。
- 保存最终系数及相应的预处理参数，用于后续预测。

课堂练习 8：比较调参准则

生成 $n = 150$、$p = 50$ 的相关线性模型，前五个系数为 2，其余为零，噪声方差为 1。固定 Lasso、同一候选网格与五折划分，比较 Min 与 1-SE 两种调参规则。

- 保存五折验证误差，分别按 Min 与 1-SE 规则选择 λ，画出 CV path 并标出两个选择。
- 分别用全部数据重拟合，比较所选 λ、非零变量数与系数估计误差 $\|\widehat{\beta} - \beta^*\|_2^2$。
- 重复 30 次，汇总上述指标的均值与波动，比较两种规则对系数收缩和变量选择的影响。

课堂练习 9：比较高维估计方法

取 $n = 100$、$p = 200$，X_{ij} 独立服从 $N(0,1)$，生成 $y = X\beta^* + \varepsilon$，$\varepsilon \sim N(0, I_n)$。比较两种等信号方差设定：

$$\text{稀疏：}\ \beta_j^* = \mathbf{1}\{j \le 5\}; \qquad \text{密集：}\ \beta_j^* = \sqrt{5/p}\quad (j = 1, \dots, p).$$

- 比较最小范数最小二乘、Ridge 与 Lasso。后两者仅用开发数据五折 CV 选 λ，再用全部开发数据重拟合。
- 每种设定重复 30 次，比较系数估计误差 $\|\widehat{\beta} - \beta^*\|_2^2$ 与非零变量数；报告均值和波动。
- 哪种方法能更准确地恢复稀疏或密集的系数？若目标是识别真实变量，评价指标应如何调整？

总结

1. 高维回归问题

高维数据案例；最小二乘与中心化；
秩、不可识别与小特征值

2. Ridge 回归

二次惩罚与闭式解；thin SVD；
系数收缩与偏差

3. 稀疏模型与最佳子集选择

活动集；ℓ_0 约束；穷举与逐步选择；
MIO、IHT、ABESS 与研究拓展

4. Lasso 估计

ℓ_1 惩罚；软阈值；
次梯度与 KKT 条件

5. 数值优化与正则化路径

坐标下降；近端梯度；
系数路径与热启动

6. 调参与模型比较

泛化误差；五折 CV 与 Min、1-SE；
全数据重拟合；稀疏与密集信号