

# 线性收缩 Lasso Ridge

殷润轩

Sun Yat-sen University

数学学院（珠海） $\text{\LaTeX}$

版本：1.0

日期：2025年3月29日

## 摘要

关键词：收缩

## 1 前置数学知识

### 1.1 范数

范数是定义在向量空间上的函数，用于衡量向量的大小。在机器学习中，常用的范数包括  $L^1$  范数和  $L^2$  范数。给定  $n$  维向量  $x = (x_1, x_2, \dots, x_n)$ ，其  $L^1$  范数和  $L^2$  范数分别定义为：

$$\|x\|_1 = \sum_{i=1}^n |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \quad (1)$$

$L^1$  范数衡量向量元素绝对值之和，常用于稀疏模型； $L^2$  范数衡量向量元素平方和的平方根，常用于最小二乘模型。

$L^2$  衡量预测真值的

$L^1$  范数用作正则化项，

使一些系数变为0，实现特征稀疏化

防止过拟合

### 1.2 向量求导

差异

定理 1.1. 方向导数与梯度的关系 若函数  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  是  $C^1$  类（一阶连续可导），则方向导数满足：

$$Df(\mathbf{x})[\mathbf{v}] = \underbrace{\mathbf{v}^T \nabla f(\mathbf{x})}_{\delta},$$

$$\nabla f(\mathbf{x}) = \left( \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)^T \quad (2)$$

其中  $\mathbf{v}$  是方向向量， $\nabla f(\mathbf{x})$  是梯度。

eg. 如  $f(x) = x_1^2 + x_2^2$

$$\begin{aligned} \nabla f(\mathbf{x}) &= (2x_1, 2x_2)^T, \text{ 取 } \mathbf{v} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T \\ Df(\mathbf{x})[\mathbf{v}] &= (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \cdot (2x_1, 2x_2)^T = 2\sqrt{2}. \end{aligned} \quad (3)$$

证明. 方向导数的定义：

$$Df(\mathbf{x})[\mathbf{v}] = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h}.$$

一阶泰勒展开：对  $f(\mathbf{x} + h\mathbf{v})$  展开：

$$f(\mathbf{x} + h\mathbf{v}) \approx f(\mathbf{x}) + h\nabla f(\mathbf{x})^T \mathbf{v} + o(h). \quad (4)$$

代入定义式得：

$$Df(\mathbf{x})[\mathbf{v}] = \nabla f(\mathbf{x})^T \mathbf{v} = \mathbf{v}^T \nabla f(\mathbf{x}). \quad (5)$$

标量  
1

□

**定理 1.2.** Hessian 矩阵与梯度方向导数的关系 若函数  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  是  $C^2$  类 (二阶连续可导), 则:

$$\nabla^2 f(\mathbf{x}) \cdot \mathbf{v} = D(\nabla f(\mathbf{x}))[\mathbf{v}], \quad (6)$$

其中  $\nabla^2 f(\mathbf{x})$  是 Hessian 矩阵,  $D(\nabla f(\mathbf{x}))[\mathbf{v}]$  表示梯度沿方向  $\mathbf{v}$  的方向导数。

证明. 梯度的方向导数: 梯度  $\nabla f(\mathbf{x})$  的方向导数为:

$$D(\nabla f(\mathbf{x}))[\mathbf{v}] = \begin{bmatrix} D(\partial_{x_1} f)[\mathbf{v}] \\ \vdots \\ D(\partial_{x_n} f)[\mathbf{v}] \end{bmatrix}.$$

Hessian 矩阵的作用: 由于  $\nabla^2 f(\mathbf{x}) = \mathbf{J}(\nabla f(\mathbf{x}))$ , 直接计算得:

$$D(\nabla f(\mathbf{x}))[\mathbf{v}] = \nabla^2 f(\mathbf{x}) \cdot \mathbf{v}. \quad (8)$$

$$\mathbf{J}(f)(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (7)$$

梯度的梯度即 Hessian 矩阵

□

## ② 局部极值判断

**例 1.1.** 我们需要找到函数  $f(x) = \frac{1}{2}x^T Ax$  的梯度 (gradient) 和 Hessian 矩阵 (Hessian matrix)。这里  $x$  是一个向量,  $A$  是一个矩阵。

方向导数  $Df(x)[v]$  表示函数  $f$  在点  $x$  处沿方向  $v$  的变化率, 定义为:

$$Df(x)[v] = \lim_{t \rightarrow 0} \frac{f(x + tv) - f(x)}{t}. \quad (9)$$

根据梯度的性质, 方向导数也可以表示为:

$$Df(x)[v] = v^T \nabla f(x) \quad (10)$$

将  $f(x + tv)$  展开:

$$f(x + tv) = \frac{1}{2}(x + tv)^T A(x + tv) \quad (11)$$

展开后得到:

$$f(x + tv) = \frac{1}{2}x^T Ax + \frac{t}{2}x^T Av + \frac{t}{2}v^T Ax + \frac{t^2}{2}v^T Av \quad (12)$$

将展开后的表达式代入方向导数的定义:

$$Df(x)[v] = \lim_{t \rightarrow 0} \frac{\frac{1}{2}x^T Ax + \frac{t}{2}x^T Av + \frac{t}{2}v^T Ax + \frac{t^2}{2}v^T Av - \frac{1}{2}x^T Ax}{t} \quad (13)$$

化简后得到:

$$Df(x)[v] = \frac{1}{2}x^T Av + \frac{1}{2}v^T Ax \quad (14)$$

由于矩阵乘法的转置性质  $v^T Ax = x^T A^T v$  如果  $A$  是对称矩阵 (即  $A = A^T$ ), 则:

$$\frac{1}{2}x^T Av + \frac{1}{2}v^T Ax = v^T Ax \quad (15)$$

根据方向导数的表达式  $Df(x)[v] = v^T \nabla f(x)$ , 我们得出:

$$\nabla f(x) = Ax \quad (16)$$

通过方向导数的定义和矩阵运算的性质, 我们推导出函数  $f(x) = \frac{1}{2}x^T Ax$  的梯度为  $\nabla f(x) = Ax$ 。Hessian 矩阵通常为  $A + A^T$ , 如果  $A$  是对称矩阵, 则 Hessian 矩阵为  $2A$ 。

## 2 矩阵求导

### 2.1 矩阵求导的定义

设标量函数  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ , 其关于矩阵  $\mathbf{X} \in \mathbb{R}^{m \times n}$  的导数为:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \dots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \dots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

$$e.g. f(\mathbf{X}) = X_{11}^2 + X_{12}X_{21}$$

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \quad \frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} 2X_{11} & X_{12} \\ X_{12} & 0 \end{bmatrix} = \frac{\partial f}{\partial \mathbf{X}}$$

### 2.2 典型公式

1. 线性函数的导数:

$$A = (a_{ij})_{m \times n}.$$

$$f(\mathbf{X}) = \text{tr}(A^\top \mathbf{X}), \quad \frac{\partial f}{\partial \mathbf{X}} = A.$$

2. 二次型的导数 ( $\mathbf{X}$  对称时):

$$f(\mathbf{X}) = \text{tr}(\mathbf{X}^\top A \mathbf{X}), \quad \frac{\partial f}{\partial \mathbf{X}} = 2AX.$$

3. 行列式的导数 ( $\mathbf{X}$  可逆时):

$$f(\mathbf{X}) = \det(\mathbf{X}), \quad \frac{\partial f}{\partial \mathbf{X}} = \det(\mathbf{X}) \cdot (\mathbf{X}^{-1})^\top.$$

4. 迹的导数:

$$f(\mathbf{X}) = \text{tr}(A \mathbf{X} B), \quad \frac{\partial f}{\partial \mathbf{X}} = A^\top B^\top.$$

5. 矩阵逆的导数:

$$\frac{\partial \mathbf{X}^{-1}}{\partial \mathbf{X}} = -\mathbf{X}^{-\top} \otimes \mathbf{X}^{-\top}.$$

克罗内克积

## 3 数学总结

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}, \quad \frac{\partial f}{\partial \mathbf{x}} = \mathbf{a}$$

$$e.g. \mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a \\ c \end{bmatrix}, \quad f(\mathbf{x}) = a + 2c \quad (17)$$

当矩阵  $\mathbf{A}$  对称时

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}, \quad \frac{\partial f}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

$$e.g. \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a \\ c \end{bmatrix}, \quad f(\mathbf{x}) = a + 2c \quad (18)$$

若  $\mathbf{A}$  不对称:

$$\frac{\partial f}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x}$$

$$e.g. \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a \\ c \end{bmatrix}, \quad f(\mathbf{x}) = a + 2c \quad (19)$$

## 4 实际观测中的最小二乘问题

假设物理量  $y$  与  $n$  个自变量  $x_1, x_2, \dots, x_n$  满足线性关系:

$$y = \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n, \quad (20)$$

通过  $m$  次观测 ( $m > n$ ) 得到超定方程组。需找到系数  $\lambda_1, \dots, \lambda_n$  最小化残差平方和。

$$\min_{\lambda_1, \dots, \lambda_n} \sum_{j=1}^m \left( y^{(j)} - \sum_{i=1}^n \lambda_i x_i^{(j)} \right)^2. \quad \sum (y^{(j)} - \lambda x^{(j)})^2. \quad (21)$$

2.2.  $f(X) = \text{tr}(X^TAX)$ .  $X$  对称,  $\frac{\partial f}{\partial X} = 2AX$

Pf.  $X = (x_{ij})_{n \times n}$ ,  $A = (a_{ij})_{n \times n}$ .

$$(X^TAX)_{kl} = \sum_{i=1}^n \sum_{j=1}^n x_{ik} \cdot a_{ij} \cdot x_{jl}$$

$$\text{tr}(X^TAX) = \sum_{k=1}^n \sum_{i=1}^n \sum_{j=1}^n x_{ik} a_{ij} x_{jk}$$

$$\therefore \frac{\partial f(X)}{\partial x_{pq}} = \sum_{j=1}^n a_{pj} x_{jq} + \sum_{i=1}^n x_{iq} a_{ip} = 2 \sum_{j=1}^n a_{pj} x_{jq}$$

$$\therefore \frac{\partial f}{\partial X} = 2AX.$$

| eg.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad X = \begin{bmatrix} x & y \\ y & z \end{bmatrix}$$

$$f(X) = x^2 + 2y^2 + z^2$$

$$\frac{\partial f}{\partial X} = 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x & y \\ y & z \end{bmatrix} = 2 \begin{bmatrix} x & y \\ y & z \end{bmatrix}$$

2.3.  $X$  逆时,  $f(X) = \det(X)$ .  $\frac{\partial f}{\partial X} = \det(X) \cdot (X^{-1})^T$

$$\det(X) = \sum_{j=1}^n (-1)^{p+j} x_{pj} \det(X_{pj})$$

$j=q$  时, 求导有非0项  
去掉  $p_3, j_3$ )

$$\text{即 } \frac{\partial f}{\partial x_{pq}} = (-1)^{p+q} \det(X_{pq}) = [\text{adj } X]_{qp}.$$

(伴随矩阵的  $(q, p)$  项)

且  $\text{adj}(X) = \det(X) \cdot X^{-1}$ ,

$$\frac{\partial f}{\partial x_{pq}} = [\det(X) \cdot X^{-1}]_{qp}.$$

$$\therefore \frac{\partial f}{\partial X} = \det(X) \cdot (X^{-1})^T.$$

| eg.  $X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$$f(X) = ad - bc$$

$$\det(X) \cdot (X^{-1})^T = \frac{(ad-bc)}{ad-bc} \cdot \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\frac{\partial f}{\partial X} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

2.4.  $f(X) = \text{tr}(AXB)$ .  $\frac{\partial f}{\partial X} = A^T B^T$

Pf.  $A = (a_{ij})_{m \times p}$   $X = (x_{ij})_{p \times q}$ .  $B = (b_{ij})_{q \times n}$

$$(AXB)_{ij} = \sum_{k=1}^p \sum_{l=1}^q a_{ik} x_{kl} b_{lj}$$

$$f(X) = \sum_{i=1}^m \sum_{k=1}^p \sum_{l=1}^q a_{ik} x_{kl} b_{lj}$$

$k=s$  且  $l=t$  时非0项

$$\frac{\partial f}{\partial x_{st}} = \sum_{i=1}^m a_{is} b_{ti} = [A^T B^T]_{st}.$$

## 2.5. 克罗内克积

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

$A \otimes B$  是  
mpx nq 的分块矩阵

$$f(x) = x^T A x$$

$$\frac{\partial f}{\partial x} = (A + A^T)x :$$

$$\text{设 } x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, a = (a_{ij})_{n \times n}.$$

$$\text{eg. } A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

$$\begin{aligned} A \otimes B &= \begin{bmatrix} B & 2B \\ 3B & 4B \end{bmatrix} \\ &= \begin{bmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{bmatrix} \end{aligned}$$

$$\frac{\partial X^{-1}}{\partial x} = -X^{-T} \otimes X^{-T}$$

$$dX \cdot X^{-1} + X \cdot dX^{-1} = 0$$

$$dX^{-1} = -X^{-1} \cdot dx \cdot X^{-1}$$

引入 vec(·), 如:  $M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$

$$\text{向量化 vec}(M) = \begin{bmatrix} m_{11} \\ m_{21} \\ m_{12} \\ m_{22} \end{bmatrix}$$

$$\therefore \text{vec}(dX^{-1}) = -\text{vec}(X^{-1} \cdot dx \cdot X^{-1})$$

$$\text{且 } \text{vec}(ABC) = (C^T \otimes A) \text{vec}(B)$$

$$\downarrow \quad \text{vec}(dx^{-1}) = -(X^{-T} \otimes X^{-T}) \text{vec}(dx)$$

若  $y = f(x)$ ,  $\frac{\partial y}{\partial x}$  有  $dy = -\text{vec}(dy)^T \cdot \text{vec}(dx)$   
对于  $y = x^T$

$$\text{vec}(dx^T) = -(X^{-T} \otimes X^{-T}) \cdot \text{vec}(dx)$$

$$\therefore \frac{\partial x^T}{\partial x} = -X^{-T} \otimes X^{-T}.$$

$$x^T A x = [x_1 \dots x_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$Ax = \begin{bmatrix} \sum_{j=1}^n a_{1j} x_j \\ \vdots \\ \sum_{j=1}^n a_{nj} x_j \end{bmatrix}$$

$$x^T A x = \sum_{i=1}^n x_i \sum_{j=1}^n a_{ij} x_j$$

$$= \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j$$

$$\frac{\partial f}{\partial x_k} = \sum_{j=1}^n a_{kj} x_j + \sum_{i=1}^n a_{ik} x_i$$

$$\frac{\partial f(x)}{\partial x} = Ax + A^T x$$

$$\text{目标: } \min_{\lambda_1, \dots, \lambda_n} \sum_{i=1}^m (y^{(i)} - \sum_{j=1}^n \lambda_j x_i^{(j)})^2$$

设计矩阵  $X$ 、观测向量  $y$ 、系数向量  $\lambda$  定义如下:

$$X = \begin{bmatrix} x_1^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(m)} & \dots & x_n^{(m)} \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}.$$

## 4.1 目标函数与解析解

在线性回归中，目标是最小化预测值与实际值之间的平方误差。这可以通过最小二乘法来实现，其数学表达式如下：

$$\min_{\lambda} \|y - X\lambda\|_2^2 = (y - X\lambda)^T (y - X\lambda) = y^T y - 2\lambda^T X^T y + \lambda^T X^T X \lambda \quad (22)$$

其中， $y$  是观测值向量， $X$  是设计矩阵（包含自变量）， $\lambda$  是回归系数向量。为了找到使目标函数最小的  $\lambda$ ，我们对目标函数关于  $\lambda$  求导并令其等于零。目标函数的导数为：

$$\nabla_{\lambda} \|y - X\lambda\|_2^2 = -2X^T(y - X\lambda) \quad \nabla_{\lambda} \|y - X\lambda\|_2^2 = -2X^T y + 2X^T X \lambda \quad (23)$$

$$= -2X^T(y - X\lambda) \quad (24)$$

将其设为零向量：

$$-2X^T(y - X\lambda) = 0$$

化简得到正规方程 (Normal Equation):

$$X^T X \lambda = X^T y \quad (25)$$

如果  $X^T X$  是可逆的，那么可以直接解出  $\lambda$ :

$$\lambda = (X^T X)^{-1} X^T y \quad (26)$$

这就是线性回归的最小二乘解。它表示在给定设计矩阵  $X$  和观测值  $y$  的情况下，使得预测误差平方和最小的回归系数  $\lambda$ 。

## 4.2 梯度下降法求数值解

梯度下降法是一种迭代优化算法，用于寻找目标函数的最小值。其基本思想是沿着目标函数梯度的反方向更新模型参数，逐步逼近最优解。目标函数为：

$$J(\lambda) = \frac{1}{2} \|y - X\lambda\|_2^2 \quad (27)$$

梯度为：

$$\nabla J(\lambda) = -X^T(y - X\lambda) \quad (28)$$

初始化回归系数向量  $\lambda$ ，然后在每次迭代中按照以下规则更新：

$$\lambda^{(k+1)} = \lambda^{(k)} - \eta \nabla J(\lambda^{(k)})$$

其中， $\eta$  是学习率，控制每次更新的步长。以下是算法流程：

1. 初始化回归系数向量  $\lambda$  为零向量或随机小值。
2. 计算当前  $\lambda$  对应的目标函数梯度  $\nabla J(\lambda)$ 。
3. 按照学习率  $\eta$  更新  $\lambda$ 。
4. 重复步骤 2 和 3，直到满足收敛条件（如梯度范数小于某个阈值或目标函数变化小于某个阈值）。

$\eta$  过大 跳过最优解而不收敛  
过小 收敛速度慢

学习率  $\eta$  是梯度下降法中的一个重要超参数。学习率过大可能导致算法发散，过小则可能导致收敛速度过慢。通常通过试验或使用学习率调度策略来选择合适的学习率。梯度下降法在目标函数为凸函数时能够保证收敛到全局最优解。对于线性回归的最小二乘问题，目标函数是凸二次函数，因此梯度下降法能够可靠地收敛到最优解。

代码在 code 中已经实现。

## 5 带罚项的模型理论与性质

不稳定 受扰动大

当矩阵  $X^T X$  不可逆或其行列式接近于零时，会导致以下问题和后果：

### 5.1 矩阵不可逆的原因

矩阵  $X^T X$  的行列式为零时，矩阵是奇异的（不可逆），这通常发生在以下情况：

- 自变量之间存在多重共线性（特征之间高度相关）。如房屋数量与可居住面积
- 自变量的数量多于样本量。
- 矩阵的列向量线性相关。房间面积(平方米)和房屋面积(平方英尺)

### 5.2 不可逆带来的问题

在普通最小二乘法 (OLS) 中，参数估计公式为：

$$\lambda = (X^T X)^{-1} X^T y \quad (30)$$

需要矩阵  $X^T X$  可逆。如果矩阵不可逆，参数无法直接计算。此外，行列式接近零时，矩阵的条件数较大，导致数值计算不稳定，可能引发误差放大。

$(X^T X + \lambda I)^{-1} X^T y$  — 岭回归

### 5.3 引入惩罚项的作用

为了处理矩阵不可逆或行列式接近零的问题，可以引入惩罚项（如岭回归中的 L2 正则化）。惩罚项通过在损失函数中添加正则化项，使得矩阵  $X^T X$  变为：

$$X^T X + \lambda I \xrightarrow{\text{正定}} \text{回归方程 } y = B_0 + B_1 x_1 + B_2 x_2 + \dots \quad (31)$$

其中  $\lambda$  是正则化参数， $I$  是单位矩阵。这种方法确保了矩阵  $X^T X + \lambda I$  是正定的，行列式大于零，从而可逆。

### 5.4 惩罚项的好处

引入惩罚项有以下好处：

- 提高数值稳定性：通过增加正则化项，矩阵的行列式增大，避免了因行列式接近零导致的数值计算问题。
- 防止过拟合：惩罚项限制了模型参数的大小，减少了模型的复杂度，从而提高模型的泛化能力。

## 6 Ridge 回归的数学原理

### 6.1 从最小二乘法开始

在最小二乘法中，参数估计公式为：

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (32)$$

其中， $X$  是设计矩阵， $Y$  是响应变量向量。当  $X^T X$  不可逆或其行列式接近于零时，会导致以下问题：

- 参数估计无法直接计算。
- 数值计算不稳定，误差可能被放大。

特征值与行列式的关系

设  $\lambda_1, \lambda_2, \dots, \lambda_p$  是  $X^T X$  的全部特征值，则：

$$|X^T X| = \prod_{i=1}^p \lambda_i \quad (33)$$

当  $|X^T X| = 0$  时，存在至少一个特征值  $\lambda_i = 0$ ，即矩阵  $X^T X$  不可逆。为了解决  $X^T X$  不可逆的问题，岭回归在损失函数中引入了惩罚项，即 L2 正则化。岭回归的参数估计公式为：

$$\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y \quad (34)$$

其中， $k$  是正则化参数， $I$  是单位矩阵。

**定理 6.1.** 设  $\lambda_1, \lambda_2, \dots, \lambda_p$  为  $X^T X$  的特征值，则  $\lambda_1 + k, \lambda_2 + k, \dots, \lambda_p + k$  为  $X^T X + kI$  的全部特征值。

**证明.** 首先，证明  $X^T X$  是半正定矩阵 (s.p.d)：

$$\forall y \in \mathbb{R}^{p \times 1}, \quad y^T X^T X y = (Xy)^T (Xy) = \langle Xy, Xy \rangle \geq 0$$

由内积的正定性可知， $X^T X$  是半正定矩阵。  
若  $Ax = \lambda x$ ，则：  
 $\lambda + k$  是  $(A + kI)x = Ax + kx = (\lambda + k)x$  的特征值。

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

半正  $\rightarrow$  正  
 $+kI$

因此， $\lambda + k$  是  $A + kI$  的特征值。  $\square$

### 7 $X^T X + kI$ 与正则化项等价的数学推导

当  $|X^T X| = 0$  或  $|X^T X| \approx 0$  时，参数估计  $\hat{\beta}$  不稳定。为了使  $\hat{\beta}$  波动较小，需要对  $\beta$  加入限制。这里将  $\hat{\beta}$  记成  $k$  的函数

$$\hat{\beta} = \hat{\beta}(k) \quad \text{—参数估计值会随 } k \text{ 变化而变化} \quad (35)$$

岭回归的目标函数定义为：

$$\hat{\beta} = \arg \min_k \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (36)$$

其中， $\lambda \sum_{j=1}^p \beta_j^2$  是惩罚项。

$\arg \min$  表示使函数取最小的自变量的取值

$$\begin{aligned} & (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \\ &= (y^T - \beta^T X^T) (y - X\beta) + \lambda \beta^T \beta \\ &= y^T y - y^T X\beta - \beta^T X^T y + \lambda \beta^T \beta \\ &\downarrow \quad + \beta^T X^T X\beta + \lambda \beta^T \beta \end{aligned}$$

$$\begin{aligned}
 & (X^T X + kI)^{-1} X^T Y \\
 &= \frac{2y^T X}{2\lambda I + 2X^T X} + \frac{2\lambda \beta}{\lambda^2} \\
 &= 0 \\
 & A_k^{-1} X^T \beta = y^T X \\
 & \beta = \frac{1}{\lambda^2 + 2X^T X} \cdot y^T X
 \end{aligned}$$

岭回归的参数估计公式为：

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

其中， $k$  是与  $\beta$  的稳定性相关联的参数。

额外说明：- 通过引入  $kI$ ，矩阵  $X^T X + kI$  的行列式增大，确保矩阵可逆。-  $k$  的选择通常通过岭迹图

#### 4. 注意事项

这里的稳定性与 Lasso 中的稀疏性是两个不同的概念。

额外说明：- 岭回归（Ridge Regression）通过 L2 正则化提高模型的稳定性。- Lasso 回归通过 L1 正则化实现稀疏性，即自动特征选择。

#### 5. 定理及其证明

定理 7.1. 目标函数可以表示为：

$$f(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (38)$$

$$\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y \quad (39)$$

证明. 原式子改写为

$$f(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \quad (40)$$

对  $\beta$  进行向量求导：

$$\nabla f(\beta) = -2X^T(y - X\beta) + 2\lambda\beta = 0 \quad (41)$$

即：

$$X^T y = (X^T X + \lambda I)\beta \quad (42)$$

解得：

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \quad (43)$$

即  $\hat{\beta} = \hat{\beta}(k)$ 。

为了说明取极小，要求二阶导数：

$$\frac{\partial^2 f(\beta)}{\partial \beta^2} = X^T X + \lambda I \quad (\text{这里省略系数2}) \quad (44)$$

要求正定。

①由一阶导得出

②Hessian ATAT 得出



□

额外说明：- 二阶导数矩阵  $X^T X + \lambda I$  的正定性确保了目标函数的凸性，从而保证解的唯一性。-  $\lambda$  的引入不仅解决了矩阵不可逆的问题，还提高了模型的数值稳定性和泛化能力。

## 8 Ridge 岭回归的收缩性质

定理 8.1. 岭回归的参数估计可以表示为：

$$\hat{\beta}(k) = A_k \beta \quad (45)$$

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

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

其中,

$$A_k = (X^T X + kI)^{-1} X^T X \quad (46)$$

证明. 岭回归的参数估计公式为:

$$\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y \quad (47)$$

普通最小二乘法的参数估计公式为:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (48)$$

将岭回归的参数估计公式表示为:

$$\hat{\beta}(k) = (X^T X + kI)^{-1} [(X^T X) \cdot (X^T X)^{-1}] X^T Y \quad (49)$$

通过分解可以得到:

$$\hat{\beta}(k) = A_k \hat{\beta} \quad (50)$$

其中,

$$A_k = (X^T X + kI)^{-1} X^T X \quad (51)$$

通过调整矩阵对普通二乘法的结果进行调整

**定理 8.2.** 奇异值分解 (SVD): 对于任意矩阵  $X \in \mathbb{R}^{m \times m}$ , 存在正交矩阵  $U$  和  $V$ , 使得:

$$X = U \Sigma V^T \quad (52)$$

其中: -  $U$  和  $V$  是正交矩阵, 满足  $U^T U = I$  和  $V^T V = I$ 。 -  $\Sigma$  是对角矩阵, 其对角线元素为奇

异值  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$ 。

具体形式为:

$$X = [u_1, u_2, \dots, u_m] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_m^T \end{bmatrix}$$

若存在  $\sigma_m = 0$ , 则对应的  $v_m$  和  $v_m^T$  不起作用。

**定理 8.3.** 对于任意  $k > 0$ , 有:

$$\|\hat{\beta}(k)\| < \|\hat{\beta}\| \quad (53)$$

其中,  $\|\cdot\|$  表示 L2 范数。

表明岭回归参数估计向量小于普通二乘法 - 长度

证明. 岭回归的参数估计公式为:

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

普通最小二乘法的参数估计公式为:

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

对矩阵  $X$  进行奇异值分解 (SVD):

$$X = U \Sigma V^T \quad (56)$$

其中,  $U$  和  $V$  是正交矩阵,  $\Sigma$  是对角矩阵, 其对角线元素为奇异值  $\sigma_1, \sigma_2, \dots, \sigma_m$ 。

较小的数值  $\Rightarrow$  模型复杂度降低  
防止过拟合

(55) 提高泛化能力

$$X = U \Sigma V^T$$

计算  $X^T X$ :

$$\underline{X^T X} = (V \Sigma U^T)(U \Sigma V^T) = V \Sigma^2 V^T \xrightarrow{\text{k.}} (X^T X + kI)^{-1} \xrightarrow{\text{V.}} \underline{V \Sigma^2 V^T} \quad (57)$$

计算  $(X^T X + kI)^{-1}$ :

$$(X^T X + kI)^{-1} = V (\Sigma^2 + kI)^{-1} V^T \xrightarrow{\text{V.}} \underline{= [V (\Sigma^2 + kI)^{-1} V^T]^{-1}} \quad (58)$$

计算  $(X^T X + kI)^{-1} X^T X$ :

$$\text{AK. } (X^T X + kI)^{-1} X^T X = V (\Sigma^2 + kI)^{-1} \Sigma^2 V^T \xrightarrow{\text{V.}} \underline{= (V^T)^T \cdot (\Sigma^2 + kI)^{-1} V} \quad (59)$$

计算  $\|\hat{\beta}(k)\|$ :

$$\|\hat{\beta}(k)\| = \|A_k \hat{\beta}\| \xrightarrow{\text{V.}} \frac{1}{(b_i^2 + \lambda)} \quad (60)$$

其中,  $A_k = (\Sigma^2 + kI)^{-1} \Sigma^2$

$$\begin{aligned} A_k &= V (\Sigma^2 + kI)^{-1} \Sigma^2 \\ &= V \cdot V^T \cdot (\Sigma^2 + kI)^{-1} \cdot \Sigma^2 \xrightarrow{\text{与k相同}} (\Sigma^2 + kI)^{-1} \Sigma^2 \\ &\quad \text{均为正则化参数.} \end{aligned} \quad (61)$$

化简后得到:

$$(\Sigma^2 + kI)^{-1} \Sigma^2 = \begin{pmatrix} \frac{\sigma_1^2}{\sigma_1^2 + k} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2 + k} & & \\ & & \ddots & \\ & & & \frac{\sigma_p^2}{\sigma_p^2 + k} \end{pmatrix} \xrightarrow{\text{L2.}} \frac{b_1^2}{b_1^2 + \lambda} \quad (62)$$

矩阵  $A_k$  的形式为:

$$A_k = \begin{pmatrix} \frac{\sigma_1^2}{\sigma_1^2 + k} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2 + k} & & \\ & & \ddots & \\ & & & \frac{\sigma_p^2}{\sigma_p^2 + k} \end{pmatrix} \xrightarrow{\text{L2.}} \quad (63)$$

由于  $\frac{\sigma_i^2}{\sigma_i^2 + k} < 1$ , 因此:

$$\|(\Sigma^2 + kI)^{-1} \Sigma^2\| \leq 1 \quad (64)$$

从而:

$$\|\hat{\beta}(k)\| \leq \|\hat{\beta}\| \quad (65)$$

**额外说明:** - 岭回归通过引入正则化参数  $k$ , 使得参数估计  $\hat{\beta}(k)$  的 L2 范数小于普通最小二乘法的参数估计  $\hat{\beta}$ 。- 奇异值  $\sigma_i$  越小, 对应的收缩程度越大, 即  $\frac{\sigma_i^2}{\sigma_i^2 + k}$  越小。- 这种收缩效果有助于减少参数估计的方差, 从而提高模型的泛化能力。

研究函数:

$$f(x) = \frac{x^2}{x^2 + k} = 1 - \frac{k}{x^2 + k} \quad (x \geq 0) \quad \frac{6_1^2}{6_1^2 + k} = \frac{100}{101} \quad \frac{6_2^2}{6_2^2 + k} = \frac{1}{7} \quad (66)$$

当  $x$  增大时,  $f(x)$  也增大, 表明较大的奇异值  $\sigma_i$  不容易被收缩。

$b_2$  收缩程度大

**结论:** 岭回归的收缩程度取决于奇异值  $\sigma_i$  的大小。较小的奇异值对应的参数估计会被更大幅度地收缩。 □

## 9 SVD 的几何意义

例子:  $X \in \mathbb{R}^{2 \times 2}$

考虑矩阵  $X \in \mathbb{R}^{2 \times 2}$ , 其奇异值分解为:

$$X = U\Sigma V^T \quad (67)$$

其中:

$$X = [u_1, u_2] \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} [v_1^T, v_2^T] \quad (68)$$

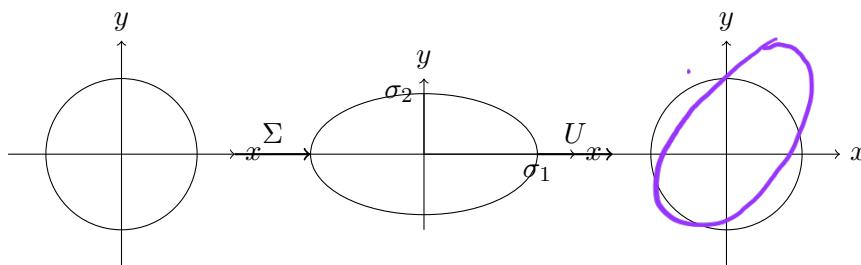


图 1: SVD 的几何意义: 单位圆经过  $\Sigma$  变换为椭圆, 再经过  $U$  变换为新的坐标系。

几何解释

- \*\*U 和 V 的作用\*\*: U 和 V 矩阵代表旋转作用。- \*\* $\Sigma$  的作用\*\*:  $\Sigma$  矩阵代表伸缩作用, 其对角线元素  $\sigma_i$  表示沿各个主轴的伸缩程度。

岭回归的几何意义

-  $\sigma_i^2$  越大, 表示该方向上的信息越多, 越不容易被收缩。- 岭回归保留信息多的方向, 剔除信息少的方向。

- $\sigma_1$  信息多, 快速收缩但不到零。
- $\sigma_2$  信息少, 快速收缩但不到零。
- $u_1$  是信息最多的方向。-  $u_2$  是信息第二多的方向。

在数据降维时, 我们会保留信息多的小方向上的特征, 对  $u_2$  进行一定程度收缩, 减小噪声影响

## 10 岭迹图分析

岭迹图 (Ridge Trace Plot) 是岭回归分析中的一种工具, 用于观察不同正则化参数  $k$  对回归系数的影响。岭迹图展示了岭回归估计量  $\hat{\beta}(k)$  的分量随  $k$  变化的轨迹。

我们可以通过解读岭迹图来判断是使用标准多元线性回归更好还是使用岭回归更好。以下是几种有代表性的情况: 如图所示, 最小二乘估计量  $\hat{\beta}$  的分量显著, 因此从标准多元线性回归的观点来看, 解释变量  $X$  的解释作用相当显著。但是  $\hat{\beta}$  的岭迹图表现出相当的不稳定性: 随着  $k$  的增大显著下降, 且趋于零的速度相当快 (尽管我们知道它一定会趋于零)。因此从岭回归的观点来看, 解释变量  $X$  的解释作用可能不显著。这说明在本例中, 标准多元线性回归可能没有反映真实情况。

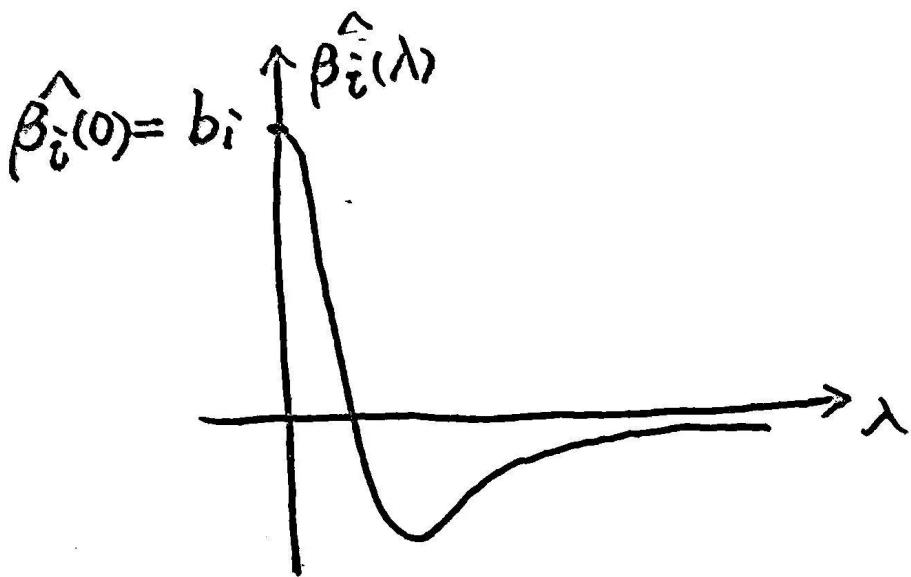


图 2: 岭迹图示例 1

### Figure 2

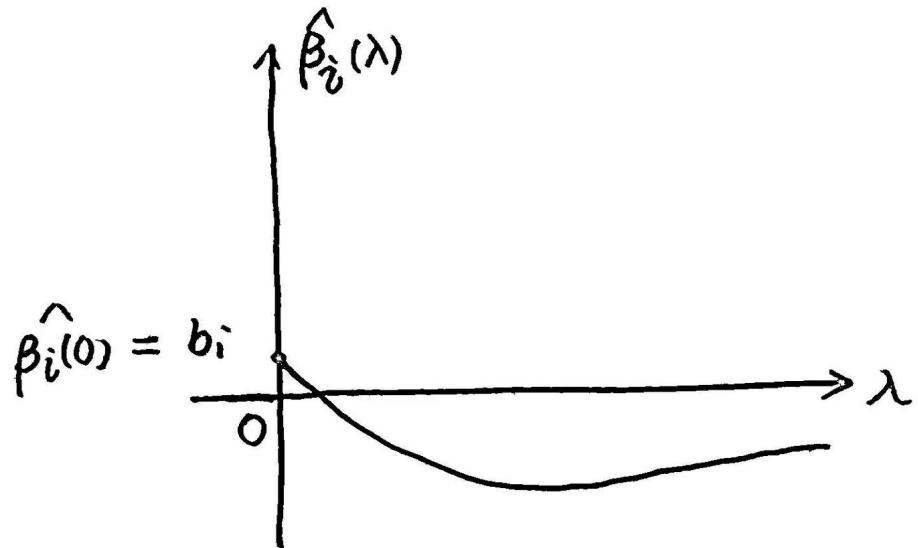
如图所示，最小二乘估计量  $\hat{\beta}$  的分量不显著，因此从标准多元线性回归的观点来看，解释变量  $X$  的解释作用不显著。但是  $\hat{\beta}$  的岭迹图显示：随着  $k$  增大，在相当长的区间内稳定为离零较远的负值（虽然我们知道它最终会趋于零）。因此从岭回归的观点来看，解释变量  $X$  的解释作用显著。这说明在本例中，标准多元线性回归没有反映真实情况。

### Figure 3

如图所示，我们发现  $\hat{\beta}_1$  和  $\hat{\beta}_2$  的岭迹曲线都很不稳定，但它们的线性组合  $\hat{\beta}_1 + \hat{\beta}_2$  却相当稳定。这说明解释变量  $X_1$  和  $X_2$  之间存在多重共线性。

### 岭迹分析的全局视角

从全局来看，岭迹分析可用于判断标准多元线性回归对某一实例是否适用。我们将所有回归系数的岭迹曲线绘制在一张图像上。如果这些岭迹曲线的不稳定性很强，即整个图像呈现比较“乱”的局面（如图 4-1 所示），那么我们倾向于认为标准多元线性回归是不适用的。此时，我们应该使用岭回归，并选取合适的岭参数  $k$ 。反之，则认为标准多元线性回归是适用的（如图 4-2 所示）。



交叉验证定量  
迹图定性

图 3: 岭迹图示例 2

岭参数  $k$  的选择

AIC / BIC 准则 — 复杂度惩罚标准, 挤惺均衡

### (1) 岭迹法 (Ridge Trace Method)

岭迹法的直观考虑是：如果标准多元线性回归的最小二乘估计量  $\hat{\beta}$  有不合理之处，那么我们希望通过采用适当的岭估计量  $\hat{\beta}(k)$  来获得一定的改善，其岭参数  $k$  的选取尤为重要。选取  $k$  的一般原则是：

- $k$  附近各回归系数  $\hat{\beta}(k)$  的岭估计量基本稳定；
- 分量的符号和绝对值与问题的 **实际意义相符**；
- $k$  与  $k$  相比不增加太多。

稳定性  
实际意义相符  
 **$k$  值适度**

具体的例子如图所示：我们发现  $k$  附近各回归系数的岭估计量基本稳定，因此我们可取  $k$ 。

值得注意的是，岭迹法与传统的基于残差的方法相比，其确定岭参数  $k$  的方法存在一定主观性，且缺乏严格的理论依据。我们可以将其作为一种定性分析，与后文所述的定量分析相辅相成。

二

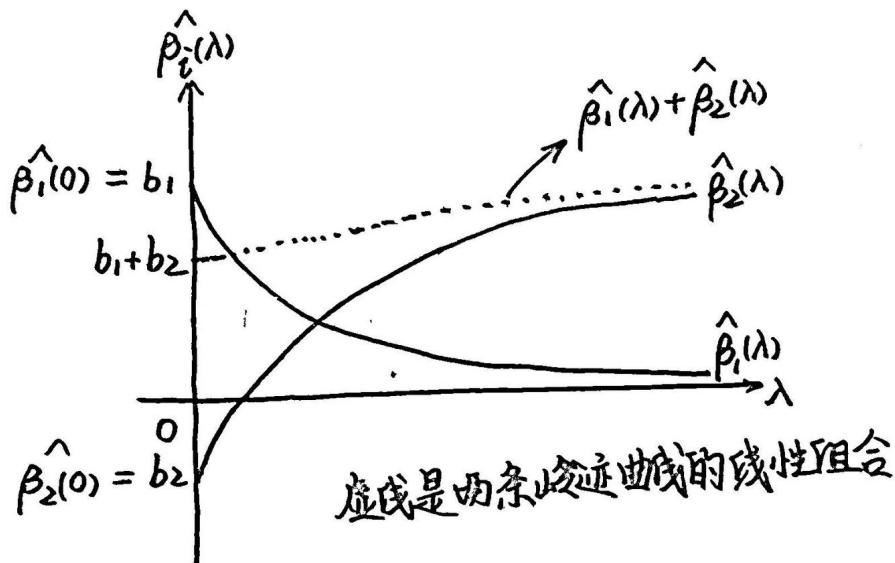


图 4: 岭迹图示例 3

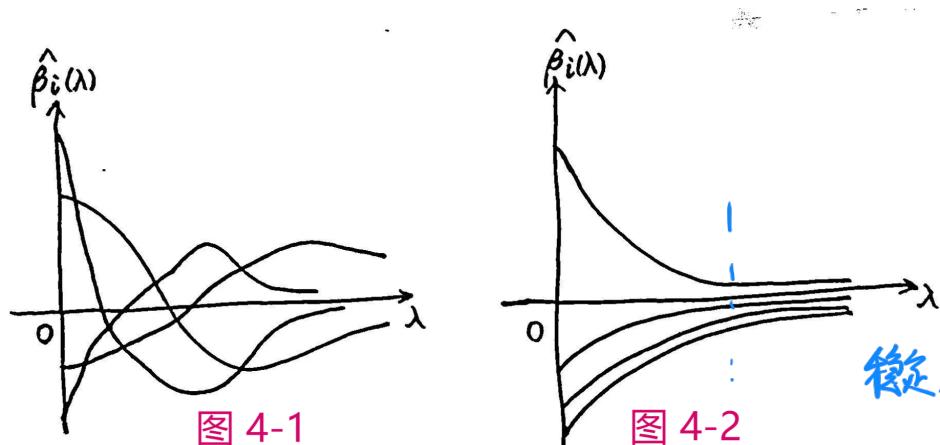


图 5: 岭迹图示例 4-1 (不适用标准多元线性回归) 4-2 (适合)

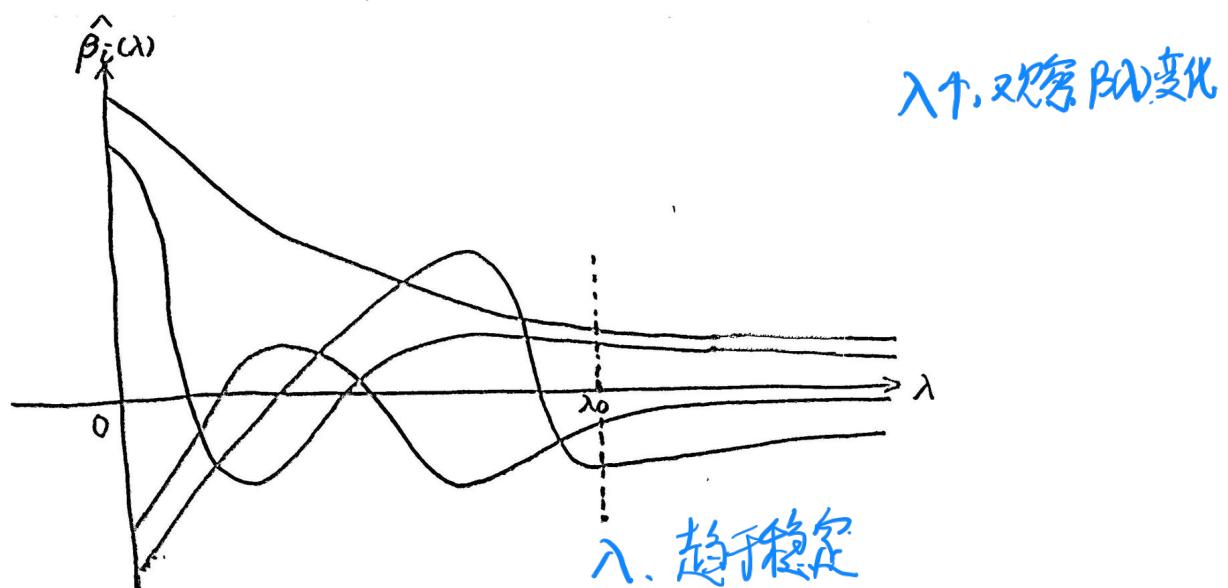


图 6: 岭迹图示例 5