中國農業大學
China Agricultural University

《机器学习》课程——

线性回归

主讲人： 徐义田教授

学　校：中国农业大学

目　录

- 问题描述
- 最小二乘回归（least square）
- 岭回归（ridge regression）
- 拉索（lasso）
- 补充材料

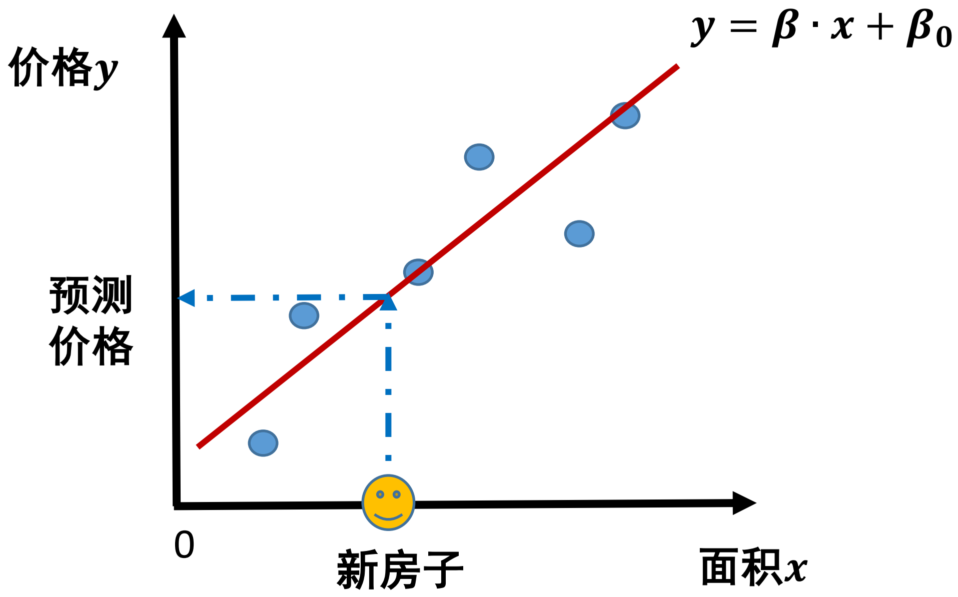
➢问题描述

一、问题描述

举例：房价预测

已知： 6所房子的面积x以及其对应的价格y

目的： 新来一所房子，预测其价格



一、问题描述

线性回归问题:

给定样本集$(x_i, y_i)_{i=1}^{n}$，其中$x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$，我们试图找到一个线性回归函数:

$$f(x) = \beta^T x + \beta_0$$

使得$f(x_i) \approx y_i$尽可能满足所有训练样本点。

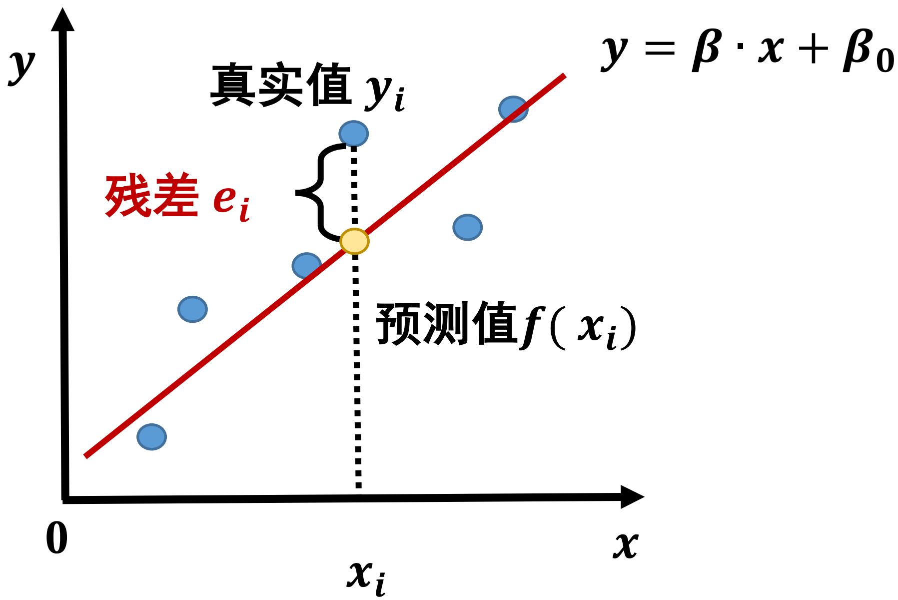
其中，β和β_0是我们需要求解的回归系数（参数）。

如何求得回归函数$f(x)$?

最小二乘回归

二、最小二乘回归

中心思想： 最优拟合直线应该使各点到直接的距离（残差）和最小，也可表述为距离的平方和最小。



$$e_i = y_i - f(x_i)$$

$$\min \sum_{i=1}^{n} {e_i}^2$$

最小化残差平方和（RSS）：$\min\limits_{\beta,\beta_0} \sum_{i=1}^{n} (y_i - \beta^T x_i - \beta_0)^2$

二、最小二乘回归

模型:

$$\min_{\beta,\beta_0} \sum_{i=1}^{n}(y_i - \beta^T x_i - \beta_0)^2$$

(矩阵形式)

$$\min_{\beta,\beta_0} RSS(\beta,\beta_0) = ||y - X\beta - \beta_0\boldsymbol{e}||_2^2$$

样本中心化去掉常数项β_0 $\quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, j = 1, \dots, p$

求解: $\quad RSS(\beta) = ||y - X\beta||_2^2$

根据最小化的一阶条件:

求导: $\frac{\partial RSS(\beta)}{\partial \beta} = -2X^T(y - X\beta) = 0$ 　得: $\boldsymbol{\beta}^* = (X^TX)^{-1}X^Ty$

二、最小二乘回归

假设回归模型形式为： $Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$,则有$Y \sim N(X\beta, \sigma^2 I)$

给定训练样本: $X = (x_1, \ldots, x_n)' \in R^{n\times m}, y \in R^{n\times 1}$

解的性质:

解: $\beta^* = (X^T X)^{-1} X^T y$

a) $\boldsymbol{E}(\beta^*) = \beta$ (无偏性)

b) $\boldsymbol{D}(\beta^*) = \boldsymbol{\sigma}^{\mathbf{2}} (X'X)^{-\mathbf{1}}$

证明:

$$\boldsymbol{E}(\beta^*) = \boldsymbol{E}((X'X)^{-1}X'y) = (X'X)^{-1}X'\boldsymbol{E}(y)$$

$$= (X'X)^{-1}X'\boldsymbol{E}(X\beta + \varepsilon) = (X'X)^{-1}X' X\beta$$

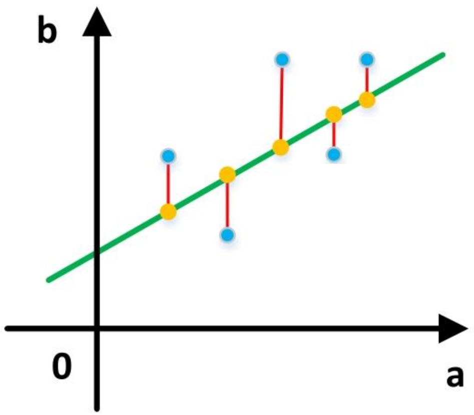
$$= \beta$$

二、最小二乘回归

$$
\begin{aligned}
\boldsymbol{D}(\beta^*) &= \boldsymbol{cov}(\beta^*, \beta^*) \\
&= \boldsymbol{E}[(\beta^* - E(\beta^*))(\beta^* - E(\beta^*))'] \\
&= \boldsymbol{E}[(\beta^* - \beta)(\beta^* - \beta)'] \\
&= \boldsymbol{E}[((X'X)^{-1}X'y - \beta)((X'X)^{-1}X'y - \beta)'] \\
&= \boldsymbol{E}[\big((X'X)^{-1}X'(X\beta + \varepsilon) - \beta\big)((X'X)^{-1}X'(X\beta + \varepsilon) - \beta)'] \\
&= \boldsymbol{E}[(\beta + (X'X)^{-1}X'\varepsilon - \beta)(\beta + (X'X)^{-1}X'\varepsilon - \beta)'] \\
&= \boldsymbol{E}[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] \\
&= (X'X)^{-1}X'\boldsymbol{E}[\varepsilon\varepsilon']X(X'X)^{-1} \\
&= (X'X)^{-1}X'\boldsymbol{E}[\sigma^2 I]X(X'X)^{-1} \\
&\boxed{= \sigma^2(X'X)^{-1}}
\end{aligned}
$$

如果$X'X$不可逆?

二、最小二乘回归



模型： $\min\ RSS(\beta) = ||y - X\beta||_2^2$

解： $\beta^* = (X^T X)^{-1} X^T y$

优点： 模型简单，且具有解析解

缺点：

1. 不能很好地处理多重共线性情况
2. 高维回归情况（$m \gg n$），过拟合，解不唯一
3. 模型不具有可解释性，即无法做特征选择

$X^T X$不可逆

在模型中加入正则化项，以限制模型复杂度，可有效缓解上述问题

注：多重共线性指样本的若干特征之间存在近似线性关系

二、最小二乘回归

模型：

$$min\underbrace{\frac{1}{2}||y-X\beta||_2^2}_{\text{损失函数}} \boxed{+\underbrace{\lambda||\beta||_p}_{\text{正则化项}}}$$

p −范数定义： $||\beta||_p$=$(\sum_{i=1}^{m}|\beta_i|^p)^{\frac{1}{p}}$

其中：

$p=2$: 岭回归,又称脊回归、吉洪诺夫正则化（Tikhonov regularization）

$p=1$: lasso

➢岭回归

（Ridge Regression）

三、岭回归

岭回归通过在 OLS 目标函数中引入2-范数正则化项，使得在严格多重共线性的情形下仍能得到唯一解，并同时起到收缩系数、缓解过拟合的作用。

模型：

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 \boxed{+ \lambda||\beta||_2^2}$$

等价形式：

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2$$

$$\boxed{s.t. \quad ||\beta||_2^2 \leq t}$$

其中，非负参数λ和t是模型的调整参数（也称“岭参数”），需人为确定其取值。

三、岭回归

China Agricultural University

模型： $\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_2^2$

求解：

由于目标函数是可微凸函数，令其一阶导数为0，得模型解析解:

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

单位矩阵的形状看起来像一条山岭，“岭回归”因此得名

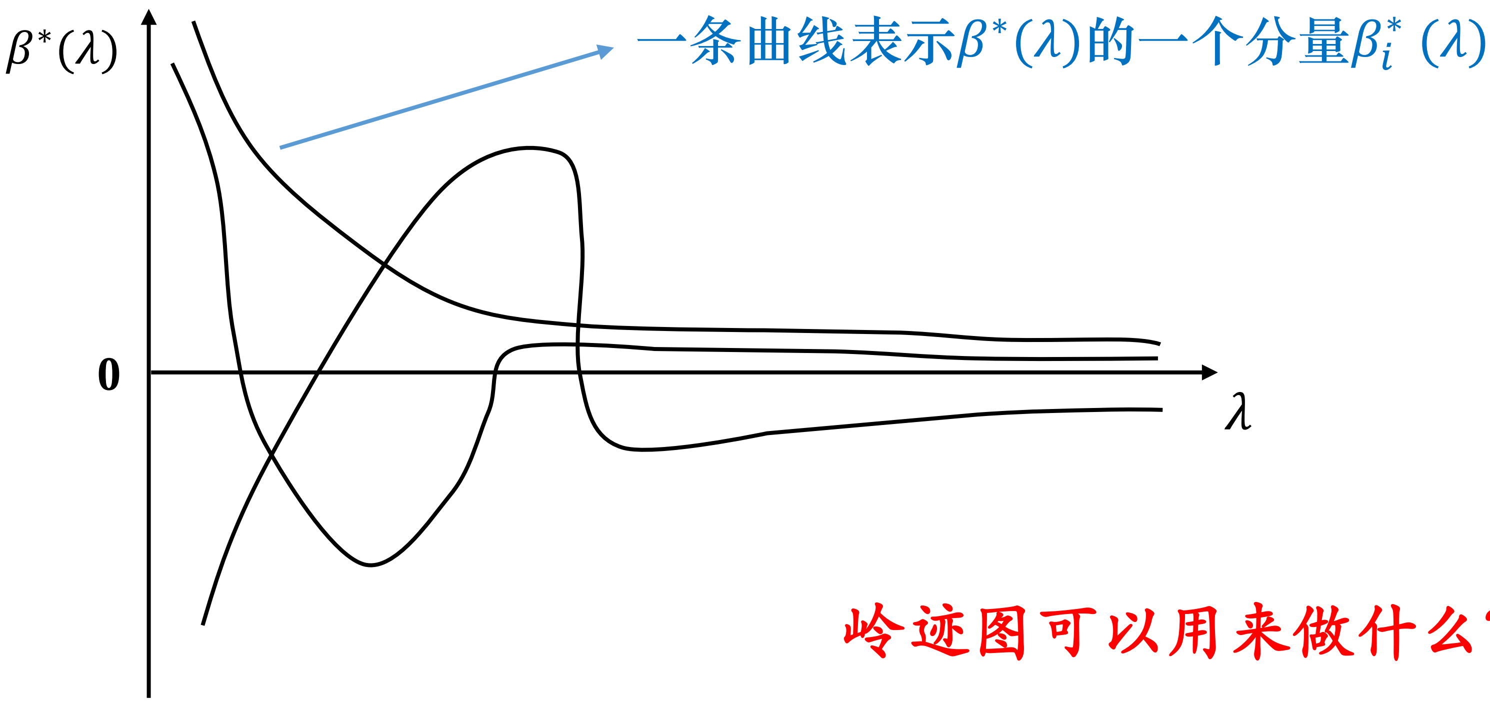
- 上述解与最小二乘相比，多了一项λI ，I为单位矩阵
- 若$X^T X$是奇异矩阵，添加λI可保证该项可逆，增强模型稳定性
- 随着λ的增大，$\beta^*(\lambda)$各元素的取值（绝对值）不断变小，最终趋于0

模型： $\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_2^2$

解： $\beta^*(\lambda) = (X^T X + \lambda I)^{-1} X^T y$

岭迹分析：

岭迹图：模型系数$\beta^*(\lambda)$随参数λ变化的曲线图



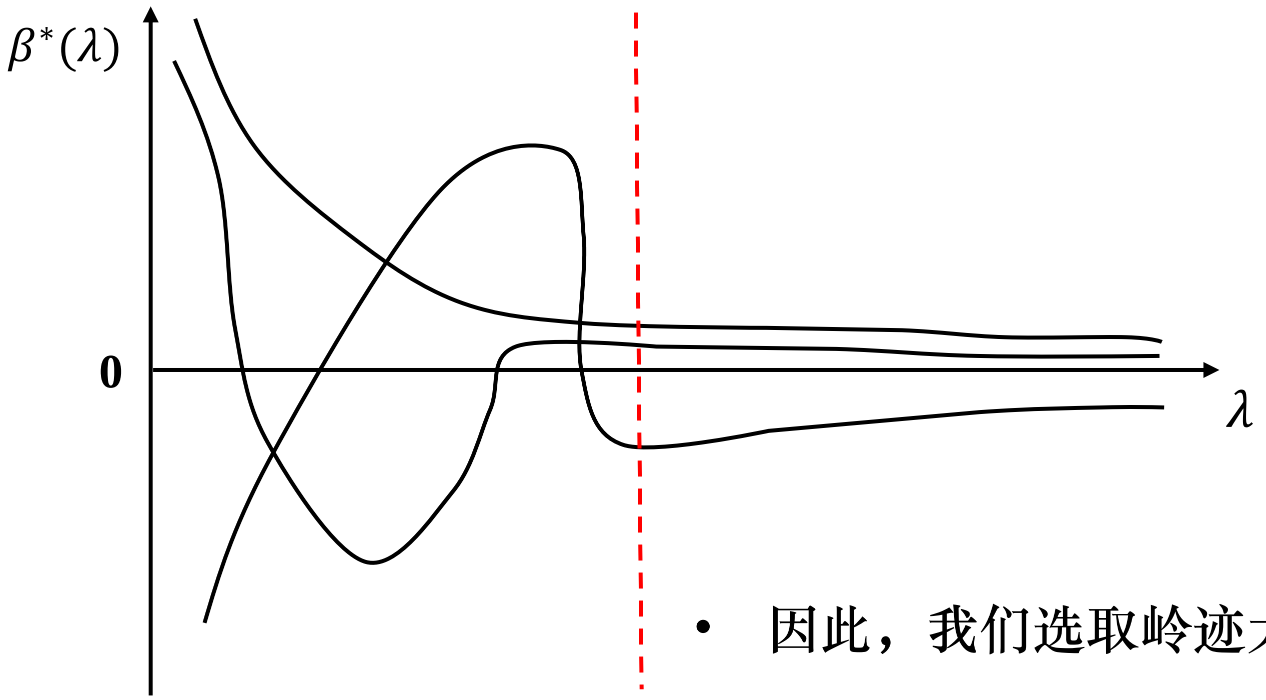
岭迹图可以用来做什么？

三、岭回归

作用一：确定适当的岭参数λ的取值

岭回归模型的解： $\beta^*(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$

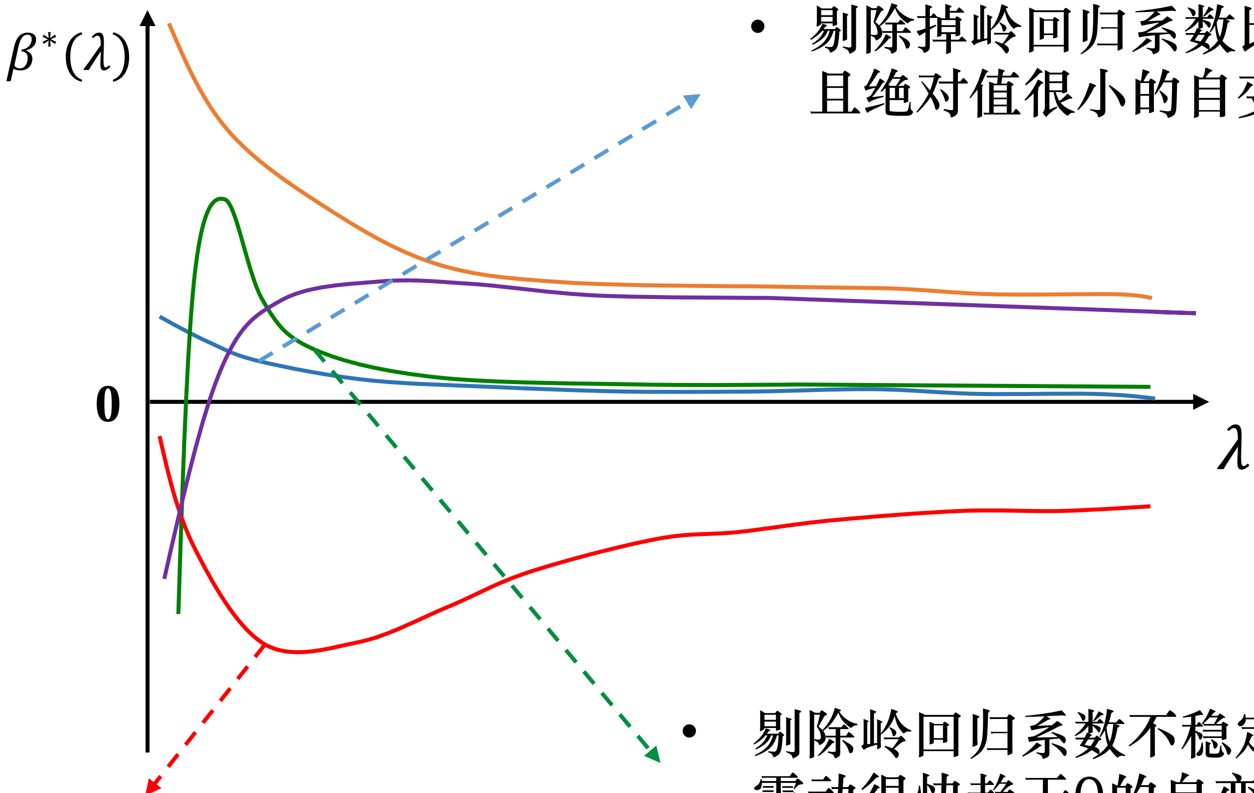
当X^TX不存在奇异性时，岭迹应是稳定地逐渐趋向于0



- 因此，我们选取岭迹大体稳定时的λ值

三、岭回归

作用二：进行自变量选择



- 剔除掉岭回归系数比较稳定且绝对值很小的自变量
- 剔除岭回归系数不稳定且震动很快趋于0的自变量
- 在最小二乘中被低估的变量

三、岭回归

模型： $\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_2^2$

解： $\beta^*(\lambda) = (X^T X + \lambda I)^{-1} X^T y$

优点：

模型中加入正则化项，解决不可逆问题，使模型更稳定。

同时缩小系数，缓解过拟合问题

缺点：

模型本身没有变量选择的作用，可解释性差

如： 如何同时考察 5 万个基因变量的回归系数?

我们通常期望从中找到真正影响疾病为数不多的基因。即我们希望真实模型是**稀疏的**

➢LASSO

1. Lasso模型

LASSO是1996年由Robert Tibshirani首次提出，全称Least absolute shrinkage and selection operator，是一种压缩估计。

模型:

$$\min_{\beta} \frac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1$$

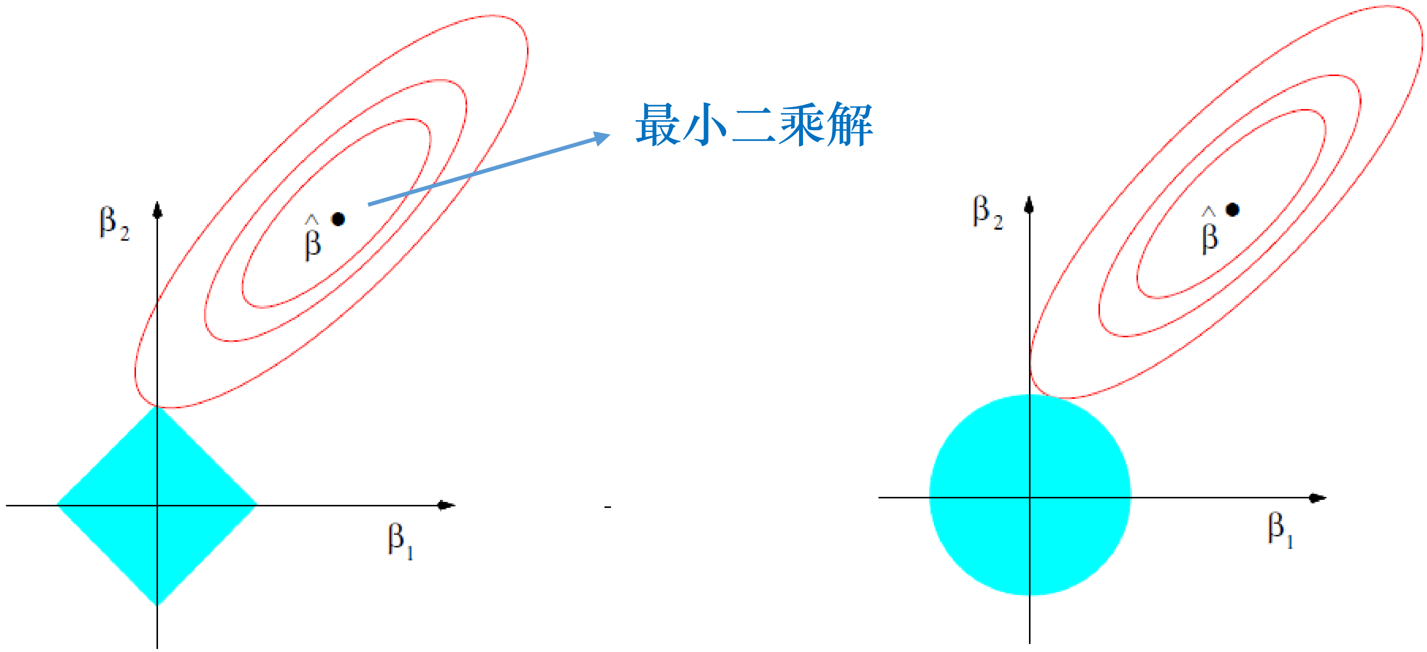
等价形式:

$$\min_{\beta} \frac{1}{2} ||y - X\beta||_2^2$$
$$s.t. ||\beta||_1 \leq t$$

与岭回归不同，lasso采用1-范数正则化项，进一步收缩系数并实现了稀疏

四、LASSO

以二维数据空间为例，从几何上说明lasso和Ridge两种方法的本质差异：



Lasso：

$$\min_{\beta_1,\beta_2} \frac{1}{2}||y - x_1\beta_1 - x_2\beta_2||_2^2$$
$$s.t. \quad |\beta_1| + |\beta_2| \le 1$$

岭回归：

$$\min_{\beta_1,\beta_2} \frac{1}{2}||y - x_1\beta_1 - x_2\beta_2||_2^2$$
$$s.t. \quad \beta_1^2 + \beta_2^2 \le 1$$

注：Lasso具有变量收缩和筛选的功能，故也称为“筛选算子”

2. Lasso的解

模型：$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$

解的唯一性（《the lasso problem and uniqueness》）

- 如果$rank(X) = m$, 解一定唯一
- 如果$rank(X) < m$, 解有可能不唯一

注：

- 对于高维回归（$m \gg n$）或特征间存在多重共线性情况，lasso的解很有可能不唯一
- 若lasso的解不唯一，则其有无数解
- 不同的解对应的模型目标函数值相等

3. 经典lasso求解算法

由于lasso模型的目标函数不连续可微（1-范数正则化项导致），因此，解没有解析形式且无法直接采用梯度下降等常规算法求解。

目前较为常用高效的lasso求解算法主要有几下几种：

① 坐标下降法（coordinate descent method）

② 近端梯度法（proximal gradient method）

③ 交替方向乘子法（ alternating direction method ）

① 坐标下降法

坐标下降法是一种非梯度优化算法。为了找到函数的局部极小值，在每次迭代中可以在当前点处沿一个坐标方向进行一维搜索。在整个过程中循环使用不同的坐标方向，直至算法收敛。

- **算法框架**

$$\min_{x\in\mathbb{R}^n} f(x)$$

分量i可按顺序更新，也可随机选取

- 给定初始值$x^{(0)}$
- 在第k（$k\geq 1$）次迭代时,已知$x^{(k-1)}$，更新：

$$x_i^{(k)} = \underset{x_i}{\operatorname{argmin}}\, f\left(x_1^{(k)}, x_2^{(k)}, \dots, x_i\ \dots, x_n^{(k-1)}\right), i \in [n]$$

- 直到：$\max_{i\in[n]} |x_i^{(k)} - x_i^{(k-1)}| \leq \varepsilon$

① 坐标下降法

Tseng（1988， 2001）证明：若目标函数$f(x)$具有如下**可分结构**，则坐标下降法可收敛到模型的全局极小值。

$$f(x) = \underbrace{g(x)}_{\text{可微凸函数}} + \sum_{j=1}^{p} \underbrace{h_j(x_j)}_{\text{凸函数}}$$

显然，lasso模型符合上述可分结构。因此，可采用坐标下降法求解lasso模型。

四、LASSO

① 坐标下降法

lasso模型： $\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda||\beta||_1$

- 数据归一化（$||x_{:i}||^2=1$），给定初始值$x^{(0)}$
- 在第k（$k\geq 1$）次迭代时,已知$x^{(k-1)}$，依次更新：

$$\boxed{x_i^{(k)}=\underset{x_i}{\operatorname{argmin}}\, f\left(x_1^{(k)},x_2^{(k)},\ldots,x_i\ \ldots,x_n^{(k-1)}\right)}, i\in[n]$$

- 直到： $\max_{i\in[n]}|x_i^{(k)}-x_i^{(k-1)}|\leq\varepsilon$

分量的更新是否具有显示表达式?

四、LASSO

中国农业大学 China Agricultural University

① 坐标下降法

$$r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\beta_k$$

$$||x_{:j}||_2^2 = 1$$

已知： $f(\beta) = \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$

求： $\beta_j^{(k)} = \underset{\beta_j}{\text{argmin}}\, f\left(\beta_1^{(k)}, \beta_2^{(k)}, \dots, \beta_j \;\dots, \beta_n^{(k-1)}\right)$

$$= \underset{\beta_i}{\text{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|$$

$$= \underset{\beta_i}{\text{argmin}} \frac{1}{2}||x_{:j}||_2^2\beta_j^2 - x_{:j}{}^T r_i^{(j)}\beta_j + \lambda|\beta_j| + const$$

求偏导得： $||x_{:j}||_2^2\beta_j - x_{:j}{}^T r_i^{(j)} + \lambda\partial|\beta_j| = 0$

其中： $\partial|\beta_j| = \begin{cases} -1, \beta_j < 0 \\ [-1,1], \beta_j = 0 \\ 1, \quad \beta_j > 0 \end{cases}$ $\Rightarrow$ $\boxed{\begin{array}{c} \beta_j^{(k)} = \mathrm{S}(x_{:j}{}^T r_i^{(j)}, \lambda) \\ S(t,\lambda) = sign(t)(|t| - \lambda)_+ \end{array}}$

② 近端梯度法

如果$f(x)$不可微，
该怎么办?

梯度下降法回顾:

考虑优化问题:

$$\min_{x} \ f(x)$$

其中，$f(x)$为可微凸函数。

则梯度下降法的迭代步骤为:

$$x^{(k+1)} = x^{(k)} - \eta \nabla f\left(x^{(k)}\right)$$

其中，η为迭代步长，可取固定值或采用线性搜索。

② 近端梯度法

近端梯度法是一种特殊的梯度下降方法，主要用于求解目标函数不可微但却可分解的最优化问题。

考虑目标函数可分解如下:

$$\min_{x} \ f(x) = g(x) + h(x)$$

$g(x)$是可微凸函数，$h(x)$是凸函数，不可微分。

近端梯度法主要包括两个步骤:

A. 对$g(x)$做二阶近似，得近端函数

B. 利用近端投影，得迭代公式

$$\min_x f(x) = g(x) + h(x)$$

② **近端梯度法**

A. 对$g(x)$做二阶近似（在已知点$x^{(k)}$处），得近端函数

$$f(x) = g(x) + h(x)$$
$$\approx g\left(x^{(k)}\right) + < \nabla g\left(x^{(k)}\right), x - x^{(k)} > + \frac{1}{2t}\left\|x - x_k\right\|_2^2 + h(x)$$
$$\approx \frac{1}{2t}\left\|x - \left(x^{(k)} - t\nabla g\left(x^{(k)}\right)\right)\right\|_2^2 + h(x) + const$$

B. 利用近端投影，得迭代公式

$$x^{(k+1)} = \underbrace{\arg\min_x \frac{1}{2t}\left\|x - \left(x^{(k)} - t\nabla g\left(x^{(k)}\right)\right)\right\|_2^2 + h(x)}$$

对于许多 ***h*** 函数，该函数有解析解

$$prox_{h,t}\left(x_k - t\nabla g\left(x^{(k)}\right)\right)$$被称为近端函数

四、LASSO

$$\min_{x}\ f(x) = g(x) + h(x)$$

② 近端梯度法

迭代公式： $x^{(k+1)} = \arg\min_{x} \frac{1}{2t}\left\|x - (x^{(k)} - t\nabla g(x^{(k)}))\right\|_2^2 + h(x)$

考虑LASSO模型：

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \qquad \begin{cases} g(\beta) = \frac{1}{2}||y - X\beta||_2^2 \\ h(\beta) = \lambda||\beta||_1 \end{cases}$$

根据上述迭代公式，可得

$$\beta^{(k+1)} = S_{\lambda t}(\beta^{(k)} - t\nabla g(\beta^{(k)}))$$

其中：$[S_{\lambda t}(z)]_i = sign(z_j)(|z_j| - \lambda t)_+$

该方法又被称作

迭代软阈值算法（ISTA）

中国农业大学 China Agricultural University

$$\min_{x} \; f(x) = g(x) + h(x)$$

② 近端梯度法

$$[S_{\lambda t}(z)]_i = sign(z_j)(|z_j| - \lambda t)_+$$

针对LASSO的迭代软阈值算法：

- 数据归一化（$||x_{:i}||^2 = 1$），给定初始值$\beta^{(0)}$
- 在第k（$k \geq 1$）次迭代时,已知$\beta^{(k-1)}$，更新：

$$\beta^{(k)} = S_{\lambda t}(\beta^{(k-1)} - tX^T(X\beta^{(k-1)} - y))$$

- 直到：$\max\limits_{i\in[n]} |x_i^{(k)} - x_i^{(k-1)}| \leq \varepsilon$

该算法的收敛速率为：$O(1/\epsilon)$，其中$\epsilon = f(x^{(k)}) - f(x^*)$

③ 交替方向乘子法

该算法常用于求解具有如下形式的优化问题：

$$\min_{x,z} f(x)+g(z)$$
$$s.t. Ax+Bz=c$$

基本步骤：

① 构造增广拉格朗日函数

$$arg\min_{x,z,\lambda} L(x,z,\lambda)=f(x)+g(z)-\lambda^T(Ax+Bz-c) + \frac{\mu}{2}||Ax+Bz-c||_2^2$$

② 交替迭代原问题和对偶问题变量

$$\begin{cases} x^{t+1}=argmin\ L(x,z^t,\lambda^t) \\ z^{t+1}=argmin\ L(x^{t+1},z,\lambda^t) \\ \lambda^{t+1}=\lambda^t-\mu(Ax^{t+1}+Bz^{t+1}-c) \end{cases}$$

③ 交替方向乘子法

lasso模型：$\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda||\beta||_1$

先将lasso模型写为ADMM标准形式：

$$\min_{x,z}\frac{1}{2}||y-X\beta||_2^2+\lambda||z||_1$$

$$s.t.\beta-z=0$$

根据ADMM算法步骤，得：

A. $$arg\min_{\beta,z,\mu}L(x,z,\lambda)=\frac{1}{2}||y-X\beta||_2^2+\lambda||z||_1-\mu^T(\beta-z)+\frac{\rho}{2}||\beta-z||_2^2$$

B. $$\begin{cases}\beta^{t+1}=(X^TX+\rho I)^{-1}(X^Ty+\mu^t+\rho z^t)\\ z^{t+1}=S_{\frac{\lambda}{\rho}}(\frac{\mu^t}{\rho}-\beta^{t+1})\\ \mu^{t+1}=\mu^t+\rho(\beta^{t+1}-z^{t+1})\end{cases}$$

China Agricultural University 中国农业大学

4. Lasso相关拓展模型

① The Elastic Net

② The Group Lasso /sparse group lasso

③ The Fused Lasso

④ Nonconvex Penalties

⑤ The Adaptive Lasso

⑥ The Bayesian Lasso

4. Lasso相关拓展模型

① The Elastic Net

问题： lasso不能很好地处理多重共线性问题

（相关特征对应的解不稳定）

但在很多实际问题中，我们希望相关性大的特征们能在问题中共同起作用/不起作用

模型：

$$\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda\left[\frac{1}{2}(1-\alpha)||\beta||_2^2+\alpha||\beta||_1\right]$$

使相关性大的特征们对应的系数相似

备注：Elastic Net增强了模型的凸性，使模型的解唯一

4. Lasso相关拓展模型

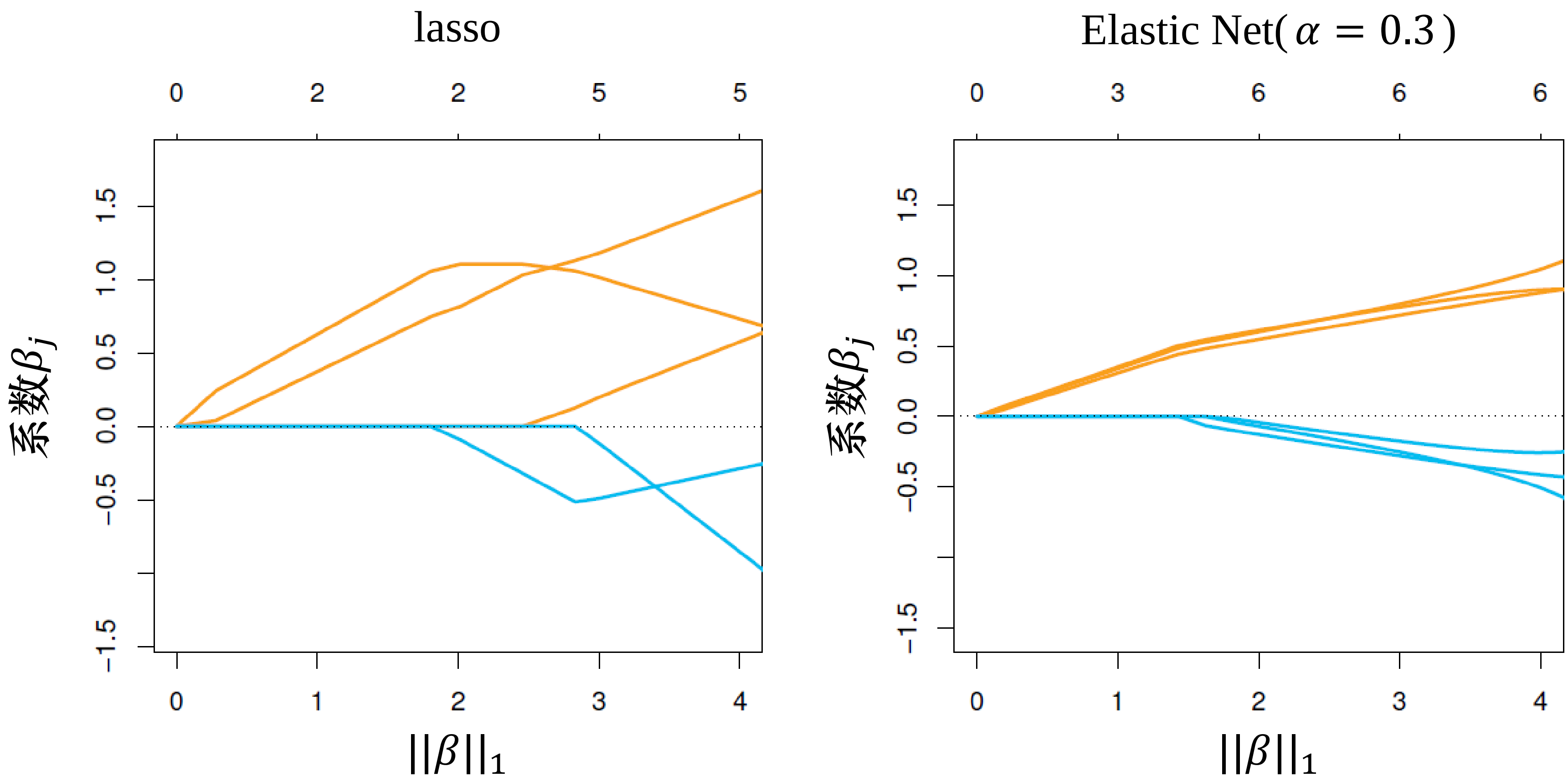
① The Elastic Net

$$Y=3Z_1-1.5Z_2+2\varepsilon, Z_1, Z_2, \varepsilon\sim N(0,1)$$

$$X_j=Z_1+\frac{\xi_j}{5}, \xi_j\sim N(0,1), j=1,2,3$$

$$X_j=Z_2+\frac{\xi_j}{5}, \xi_j\sim N(0,1), j=4,5,6$$

举例:



中国农业大学 China Agricultural University

4. Lasso相关拓展模型

② The Group Lasso

问题： 在很多实际回归问题中，特征有分组的情况

假设特征有J组（$j=1,2,\ldots,J$），$Z_j \in \mathbb{R}^{p_j}$是第j组的数据，因此，$X =$（$Z_1,\ldots,Z_J$）

模型：

$$\min_{\beta_j \in \mathbb{R}^{p_j}} \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{J} z_{i,j}\beta_j\right)^2 + \lambda \sum_{j=1}^{J}\sqrt{p_j}\,||\beta_j||_2$$

权重
避免规模大的组被留下

4. Lasso相关拓展模型

② The sparse group lasso

问题： 在很多实际问题中，我们希望在选出有用的特征组之后，还能进一步找到组内真正起作用的特征。

思想： 组稀疏 + 组内稀疏

模型：

$$\min_{\beta_j \in \mathbb{R}^{p_j}} \frac{1}{2} ||Y - \sum_{j=1}^{J} Z_j \beta_j ||_2^2 + \lambda \sum_{j=1}^{J} [(1-\alpha)||\beta_j||_2 + \alpha||\beta_j||_1]$$

4. Lasso相关拓展模型

Elastic Net 模型：

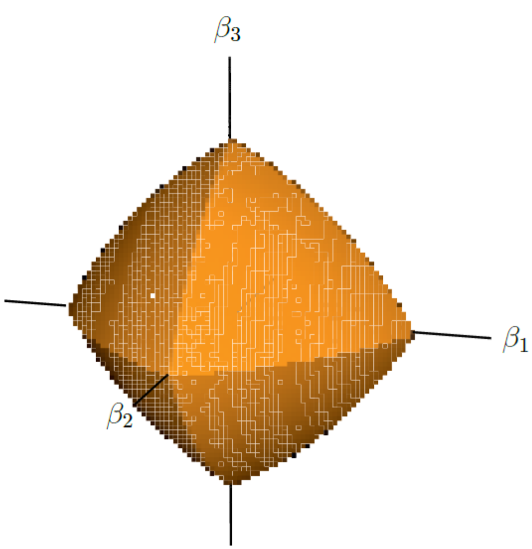
$$\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda[\frac{1}{2}(1-\alpha)||\beta||_2^2+\alpha||\beta||_1]$$

Group lasso 模型：

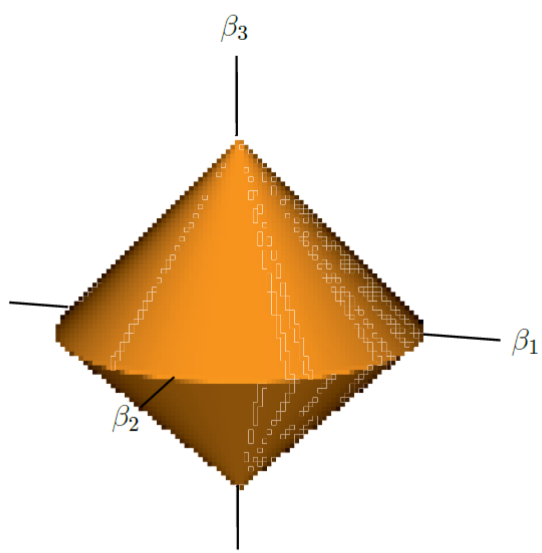
$$\min_{\beta_j\in\mathbb{R}^{p_j}}\frac{1}{2}\sum_{i=1}^{N}(y_i-\sum_{j=1}^{J}z_{i,j}\beta_j)^2+\lambda\sum_{j=1}^{J}\sqrt{p_j}||\beta_j||_2$$

Sparse Group Lasso模型：

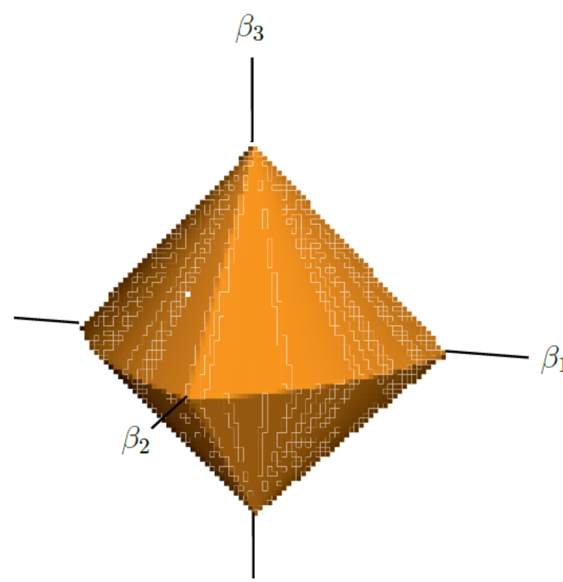
$$\min_{\beta_j\in\mathbb{R}^{p_j}}\frac{1}{2}||Y-\sum^{J}Z_j\beta_j\,||_2^2+\lambda\sum^{J}[\,(1-\alpha)||\beta_j||_2+\alpha||\beta_j||_1]$$



Elastic Net(α =0.7)



group lasso



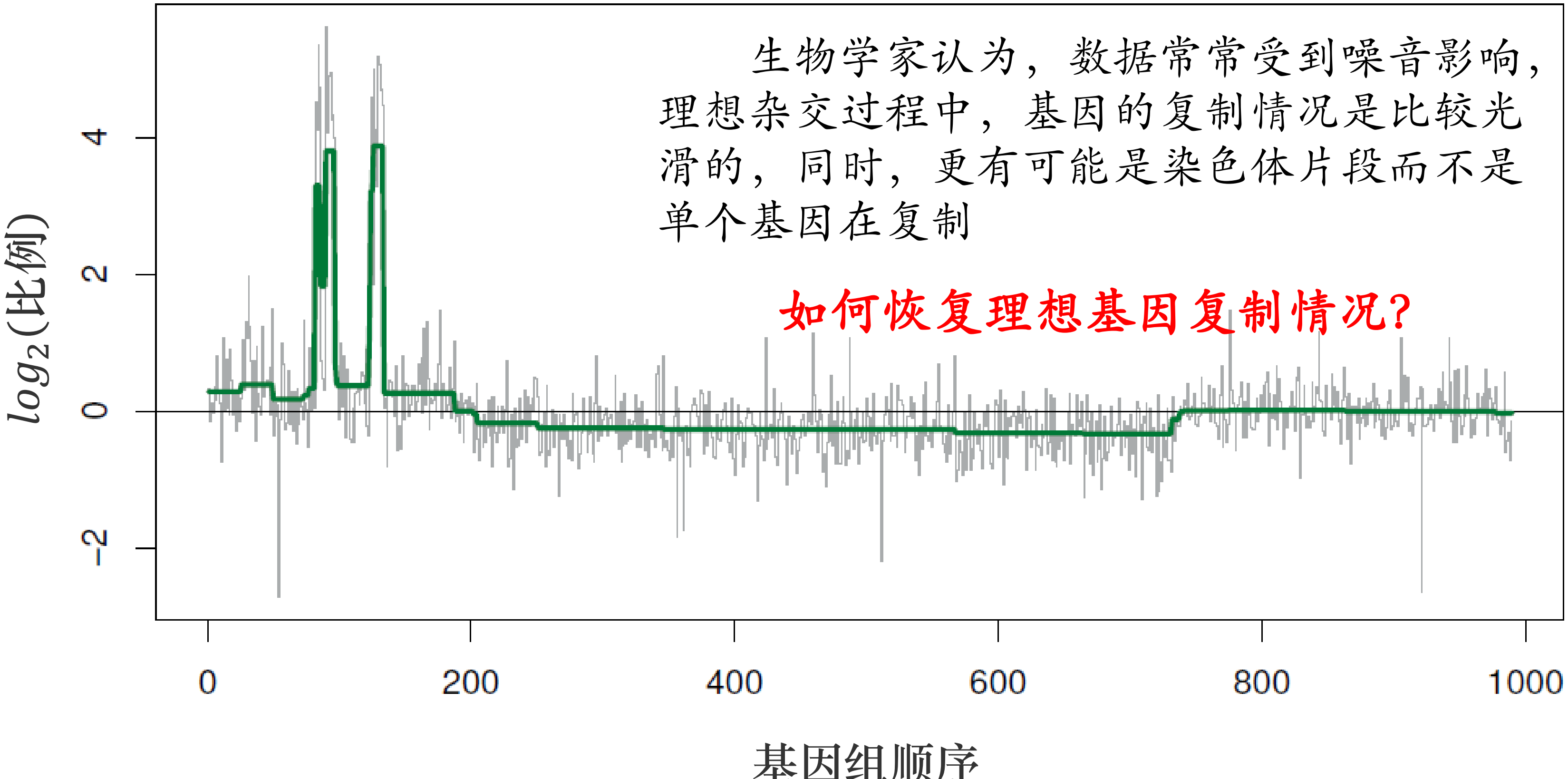
sparse group lasso(α = 0.5)

四、LASSO

4. Lasso相关拓展模型

③ The fused lasso

举例： 比较基因组杂交 (comparative genomic hybridization CGH)



4. Lasso相关拓展模型

③ The fused lasso

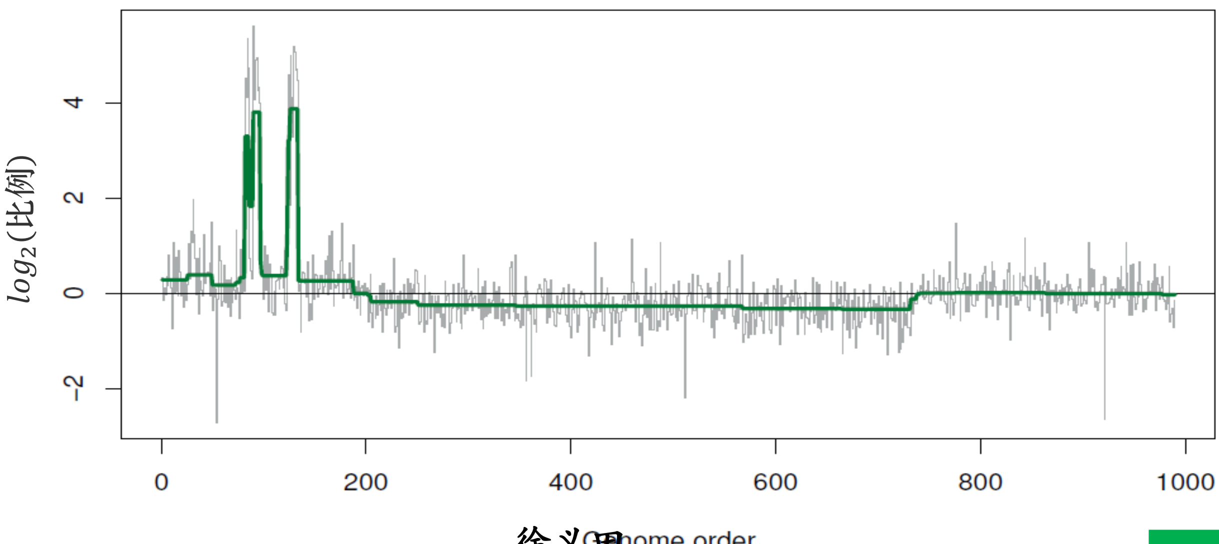
构造一个拟合函数，使结果: (1) 稀疏; (2) 光滑/piecewise-constant

$$\min_{\theta \in \mathbb{R}^N} \frac{1}{2}\sum_{i=1}^{N}(y_i - \theta_i)^2 + \lambda_1 \sum_{i=1}^{N}|\theta_i| + \boxed{\lambda_2 \sum_{i=2}^{N}|\theta_i - \theta_{i-1}|}$$

使相邻的系数取值相似/近乎相等

（total-variation denoising）

去
噪



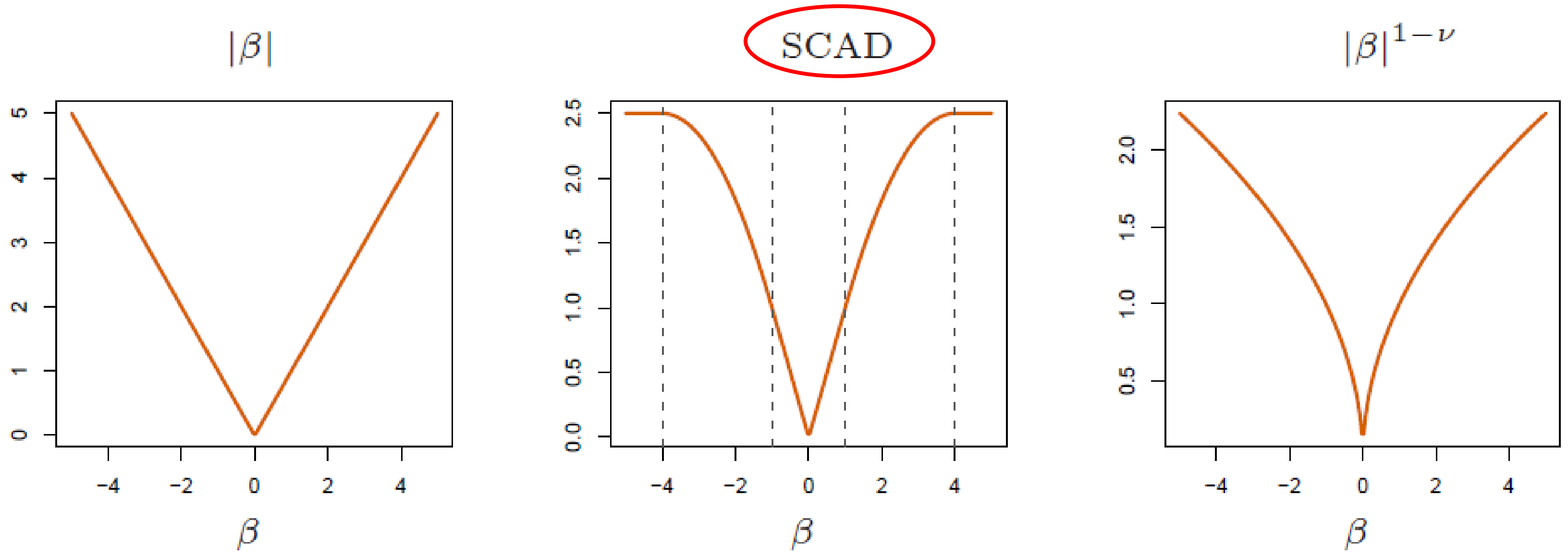
四、LASSO

4. Lasso相关拓展模型

lasso模型：$\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda||\beta||_1$

④ Nonconvex Penalties

问题：为了解决一范数为了实现稀疏性，将有些系数过于收缩（有偏）的缺点，有很多非凸惩罚项被提出



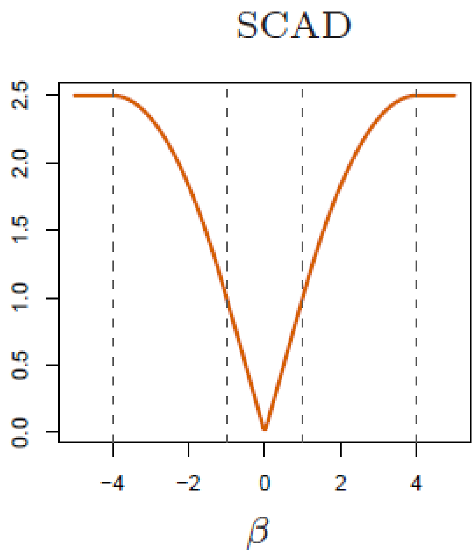
4. Lasso相关拓展模型

④ Nonconvex Penalties——SCAD penalty (Fan and Li 2005)

(smoothly clipped absolute deviation)

模型:

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



其中

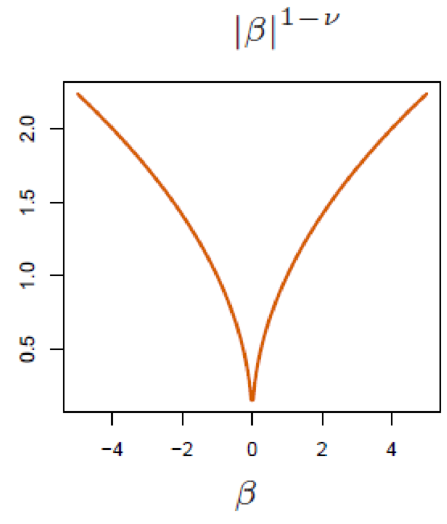
$$P_{\lambda}^{SCAD}(\beta) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda \\ -\dfrac{|\beta|^2 - 2\alpha\lambda|\beta| + \lambda^2}{2(\alpha-1)}, & \lambda < |\beta| \le \alpha\lambda \\ \dfrac{(\alpha+1)\lambda^2}{2}, & |\beta| > \alpha\lambda \end{cases}$$

4. Lasso相关拓展模型

⑤ The adaptive lasso

问题： SCAD 和$|\beta|^{1-v}$是非凸的，计算不太方便。我们希望能构造一个凸函数，具有与它们类似的优点。

$$\min_{\beta}\frac{1}{2}||y-X\beta||_2^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|$$



其中权重$w_j=1/|\hat{\beta}_j|^v$, $\hat{\beta}_j$是最小二乘估计的值。

采用这样的方法，该模型能给出与$|\beta|^{1-v}$类似的结果，但同时又保留了模型的凸性。

4. Lasso相关拓展模型

① The Elastic Net ② The Group Lasso /sparse group lasso	• 变量选择
③ The Fused Lasso	• 去噪
④ Nonconvex Penalties ⑤ The Adaptive Lasso	• 无偏性/改进预测效果
⑥ The Bayesian Lasso	• 跟概率统计结合起来

关于最小二乘、岭回归、LASSO三个模型的概率解释：

对于线性回归问题，假设模型形式为：$y = w^T x + \varepsilon$

给定训练样本: $X = (x_1, \dots, x_n)' \in R^{n\times d}, Y \in R^{n\times 1}$　　求参数：w？

- **最小二乘：**

假设 $\boldsymbol{\varepsilon \sim N(0, \sigma^2 I)}$，故有 $y \sim N(X\beta, \sigma^2 I)$。采用最大似然估计：

$$\max_w L(w)$$

$$= ln \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{y_i - w^T x_i}{\sigma}\right)^2\right)$$

$$= -\frac{1}{2\sigma^2} \boxed{\sum_{i=1}^{n} (y_i - w^T x_i)^2} - nln\sigma\sqrt{2\pi}$$

关于最小二乘、岭回归、LASSO三个模型的概率解释:

- **岭回归:**

假设 $\varepsilon \sim N(0,\sigma^2 I)$，$w_i \sim N(0,\frac{1}{\lambda}I)$。则有:

$$
\begin{aligned}
&\max_w L(w) \\
&= P(x,y|w) \times P(w)
\end{aligned}
$$

$$
= \ln\{\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{y_i - w^T x_i}{\sigma}\right)^2\right) \cdot \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi/\lambda}} exp\left(-\frac{1}{2}\left(\frac{w_i}{1/\sqrt{\lambda}}\right)^2\right)\}
$$

$$
= \boxed{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \lambda\sum_{j=1}^{d} w_j^2} + const
$$

五、补充材料

关于最小二乘、岭回归、LASSO三个模型的概率解释：

- **LASSO：**

假设 $\varepsilon \sim N(0, \sigma^2 I)$，$w_i \sim Laplace(0, b)$。 则有：

$$\max_w L(w)$$

$$= P(x, y|w) \times P(w)$$

$$= \ln\{\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{y_i - w^T x_i}{\sigma}\right)^2\right) \cdot \prod_{j=1}^{d} \frac{1}{2b} exp\left(-\frac{|w_j|}{b}\right)\}$$

$$= -\boxed{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \lambda \sum_{j=1}^{d}|w_j| +} \, const$$

谢 谢！