

5

Regularized Regression

正则化回归

利用正则项，缩减特征，构造简洁模型



遇到数学难题，别犯愁；困扰我的难题比你的大得多。

Do not worry too much about your difficulties in mathematics, I can assure you that mine are still greater.

—— 阿尔伯特·爱因斯坦 (Albert Einstein) | 理论物理学家 | 1879 ~ 1955



- ◀ `seaborn.lineplot()` 绘制线图
- ◀ `sklearn.linear_model.ElasticNet()` 求解弹性网络回归问题
- ◀ `sklearn.linear_model.lars_path()` 生成 Lasso 回归参数轨迹图
- ◀ `sklearn.linear_model.Lasso()` 求解套索回归问题
- ◀ `sklearn.linear_model.Ridge()` 求解岭回归问题
- ◀ `sklearn.metrics.mean_squared_error()` 计算均方误差 MSE
- ◀ `statsmodels.api.add_constant()` 线性回归增加一列常数 1
- ◀ `statsmodels.api.OLS()` 最小二乘法函数



5.1 正则化：抑制过度拟合

正则化 (regularization) 可以用来抑制过度拟合。本书前文提过，所谓过度拟合，是指模型参数过多或者结构过于复杂。

正则项 (regularizer, regularization term, penalty term) 通常被加在**目标函数** (objective function) 当中。正则项可以让估计参数变小甚至为 0，这一现象也叫**特征缩减** (shrinkage)。本章将采用图形方式来讲解如何在多元线性回归目标函数中引入正则项。

本章将 L1 正则项、L2 正则项以及 L1 和 L2 混合正则项利用在多变量线性回归中。L1 正则化为回归参数的 L^1 范数，L2 正则化为回归参数的 L^2 范数。

⚠ 鸢尾花书中在谈及 L^p 范数时，会采用相对严格的数学记号 L^p 。

OLS 优化问题

对于多元线性 OLS 回归，优化问题为：

$$\arg \min_b \|y - Xb\|_2^2 \quad (1)$$

对于二元线性 OLS 回归，不考虑常数项系数， b_1 和 b_2 两个回归参数形成如图 1 所示曲面。容易发现曲面为二次椭圆曲面。

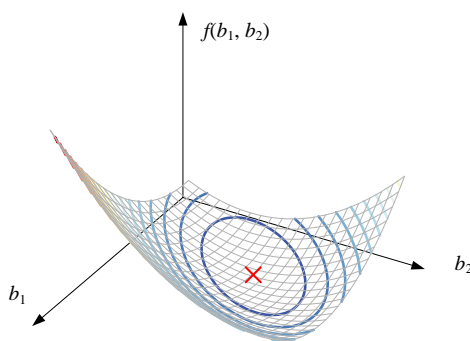


图 1. 二元线性 OLS 回归参数曲面

L2 正则化

线性 OLS 中，引入 L2 正则项，可以得到**岭回归** (ridge regression)：

$$\arg \min_b \|y - Xb\|_2^2 + \underbrace{\alpha \|b\|_2^2}_{\text{regularizer}} \quad (2)$$

白话说，L2 正则化是回归参数各个元素平方之和。 α 这个惩罚系数是用户决定的。

⚠ 注意，一般文献中上式惩罚系数用 λ ，本章和 Scikit-learn 保持一致采用 α 。

(2) 相当于图 1 曲面叠加了 L2 正则项曲面，具体如图 2。L2 正则项曲面等高线为正圆面，对应的最小值点为原点。叠加得到的岭回归参数曲面最小值位置朝原点发生明显偏移。

当 (2) 中参数 α 越大，正则项影响越大，求解优化问题得到的回归参数越靠近原点。

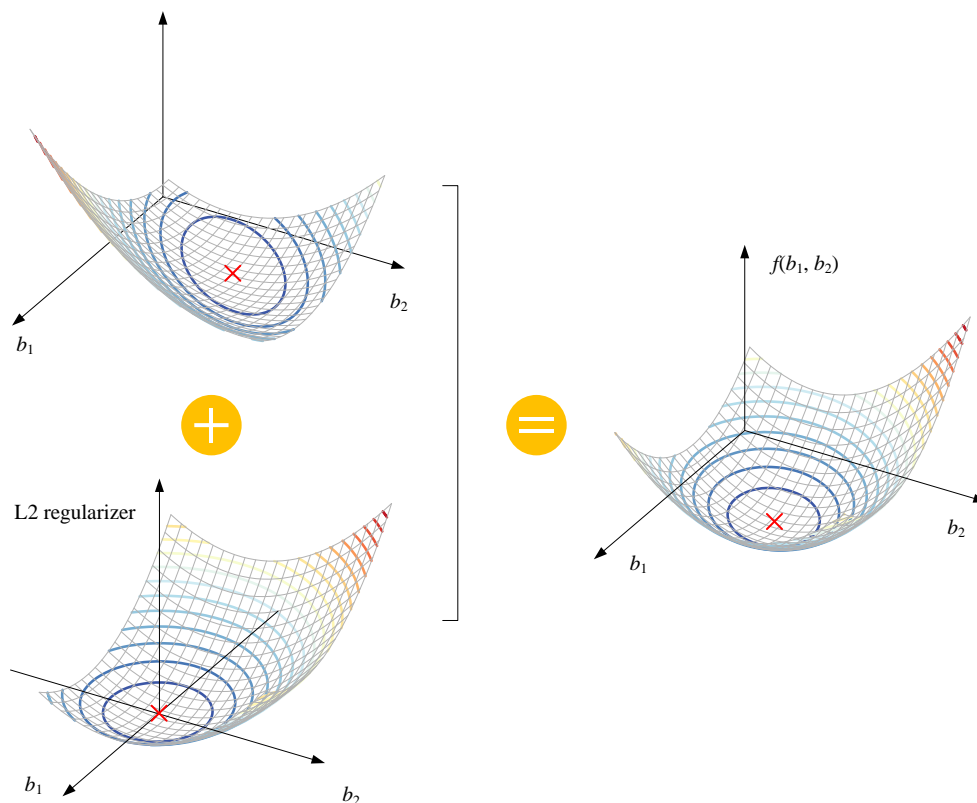


图 2. 岭回归参数曲面

L1 正则化

线性 OLS 中，引入 L1 正则项，可以得到**套索回归** (LASSO regression)：

$$\arg \min_b \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \underbrace{\alpha \|\mathbf{b}\|_1}_{\text{regularizer}} \quad (3)$$

⚠ 注意，(3) 中多元线性 OLS 回归优化项除以 $2n$ ， n 为样本数据数量。此外，不同文献套索回归的目标函数稍有不同，本章和 Scikit-learn 保持一致。

白话说，L1 正则化是回归参数各个元素绝对值之和。



鸢尾花书《矩阵力量》介绍过 L1 正则项曲面等高线为旋转正方形。

(3) 相当于在图 1 二次椭圆抛物面上叠加 L1 正则项曲面。图 3 所示为这一过程。套索回归可以进行特征选择，从而有效减少回归模型所依赖的特征数量，本章后文将从不同角度详细讲解这一点。

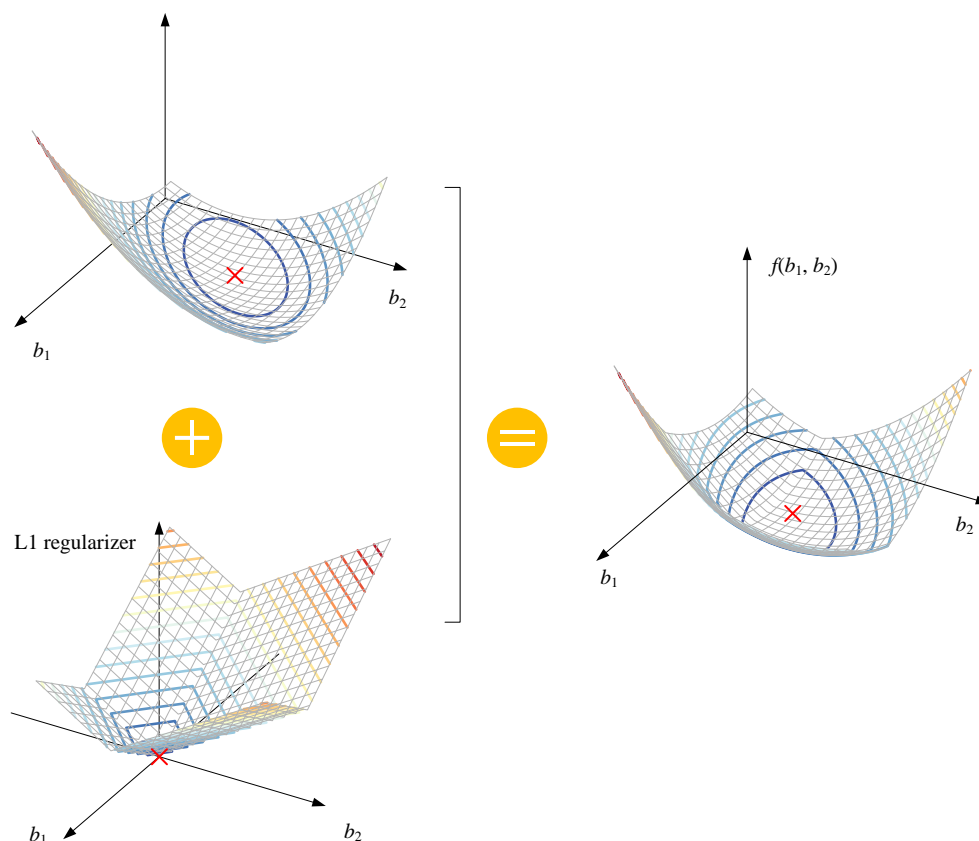


图 3. 套索回归参数曲面

L1 + L2 正则化

线性 OLS 中，以不同比例同时引入 L1 和 L2 正则项，可以得到**弹性网络回归** (elastic net regression)：

$$\arg \min_b \frac{1}{2n} \|y - Xb\|_2^2 + \alpha \left(\rho \|b\|_1 + \frac{(1-\rho)}{2} \|b\|_2^2 \right) \quad (4)$$

其中，参数 ρ 用来调和 L1 和 L2 正则项的比例。图 4 所示如何构造得到弹性网络回归系数曲面。弹性网络回归相当于岭回归和套索回归的合体。

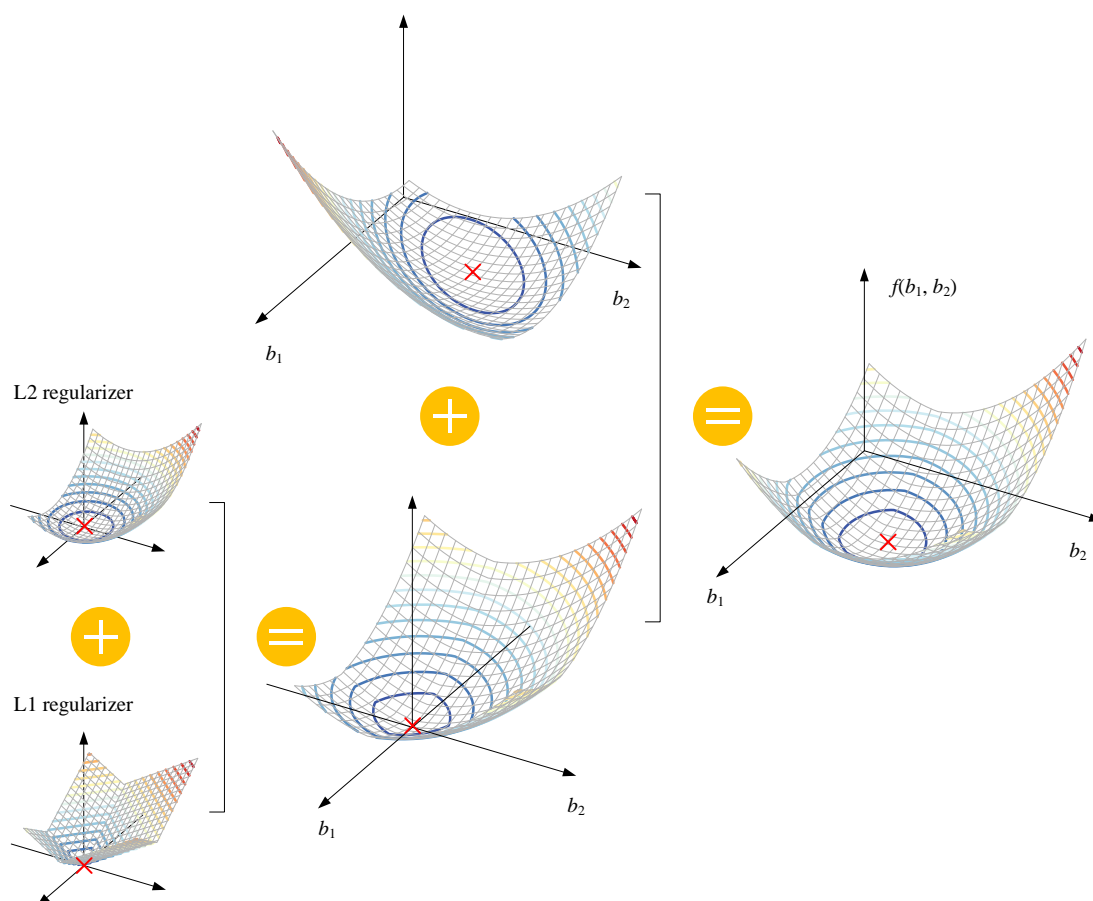


图 4. 弹性网络回归参数曲面

5.2 岭回归

如前文所述，岭回归引入 L2 正则项来缩减模型参数，岭回归的优化目标函数为：

$$f(\mathbf{b}) = \underbrace{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2}_{\text{OLS}} + \underbrace{\alpha \|\mathbf{b}\|_2^2}_{\text{L2 regularizer}} \quad (5)$$

图 5 所示为给定 α 条件下，(5) 如何构造得到岭回归目标函数参数曲面等高线图。

⚠ 注意，本节假设回归问题为二元，只有 b_1 和 b_2 两个回归参数，并且不考虑常数项系数。

如前文所述，(5) 目标函数中 OLS 部分对应椭圆抛物面，最小值点为红色 ×；红色 × 为二元 OLS 线性回归参数解的位置。

(5) 中 L2 正则项则对应正圆抛物面，最小值点为蓝色 ×，位于原点。原点处，参数系数为全 0。

➡ 根据《数学要素》一书中介绍的二次曲面内容，两个二次曲面叠加得到的一般还是一个二次曲面。

(5) 对应的曲面 $f(b_1, b_2)$ 仍然是一个椭圆抛物面，最小值点为黄色 ×；黄色 × 为给定 α 条件下岭回归参数的优化解。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

容易发现，黄色 \times 位于红色 \times 和蓝色 \times 之间；相对 OLS 线性回归参数红色 \times ，岭回归参数黄色 \times ，更靠近原点。

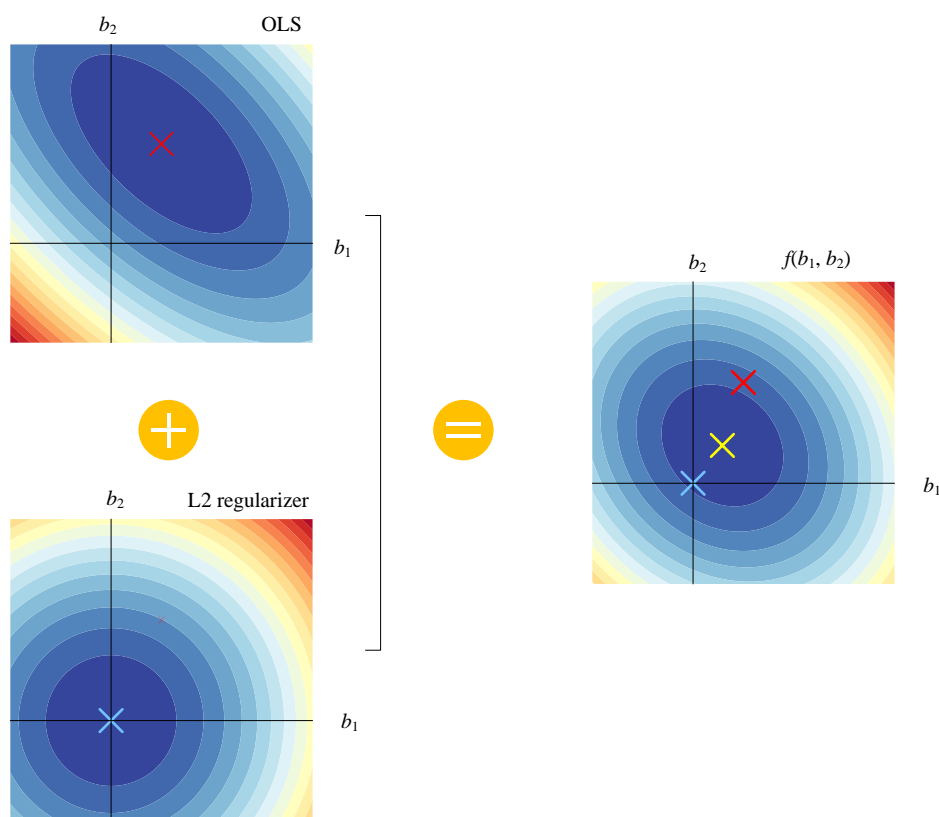
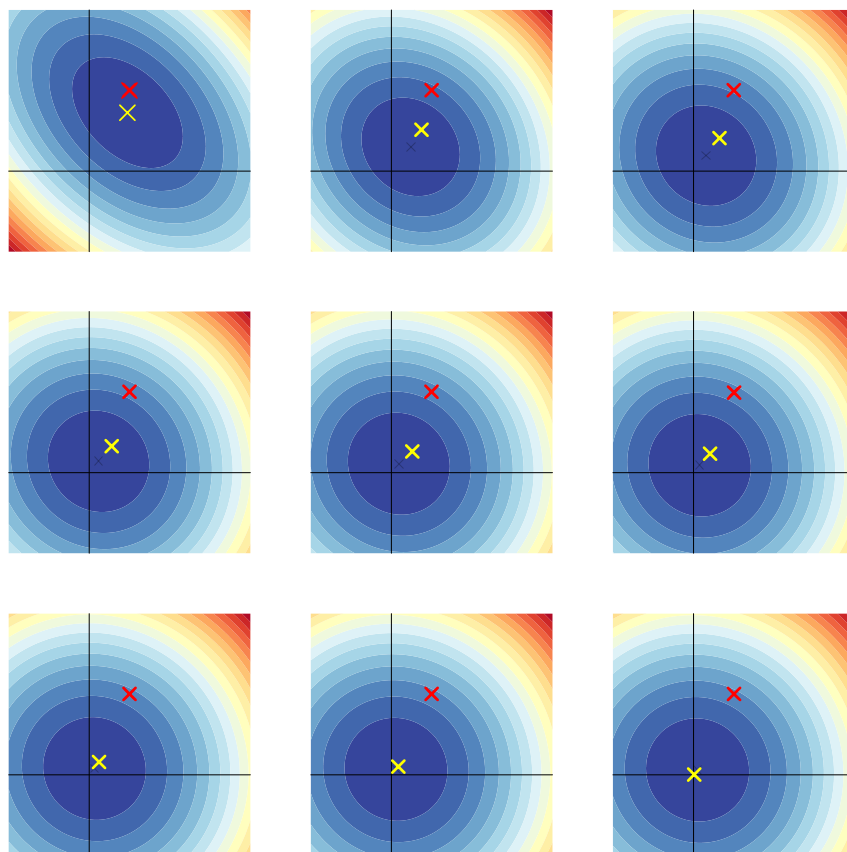


图 5. 构造岭回归优化问题参数曲面

不断增大 L2 约束项参数 α ，可以发现岭回归参数优化解不断靠近原点，如图 6 所示。注意，图 6 分图中的等高线为岭回归曲面 $f(b_1, b_2)$ 。当约束项参数 α 不断增大， $f(b_1, b_2)$ 曲面中 L2 正则项（正圆曲面）影响力不断增强。参数 α 不断增大， $f(b_1, b_2)$ 曲面等高线也从旋转椭圆渐渐变成正圆，最小值点也渐渐靠近（收缩到）原点。

图 6. 不断增大 α , 岭回归参数位置变化

构造一个线性回归问题，利用 12 只股票的日收益率解释标普 500 涨跌。图 7 所示为利用 OLS 多元线性回归得到的这个回归问题的参数。

OLS Regression Results						
Dep. Variable:	SP500	R-squared:	0.774			
Model:	OLS	Adj. R-squared:	0.750			
Method:	Least Squares	F-statistic:	32.48			
Date:	XXXXXXXXXXXXXXX	Prob (F-statistic):	3.03e-31			
Time:	XXXXXXXXXXXXXXX	Log-Likelihood:	493.88			
No. Observations:	127	AIC:	-961.8			
Df Residuals:	114	BIC:	-924.8			
Df Model:	12					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.0005	0.000	-1.038	0.302	-0.001	0.000
TSLA	0.0248	0.011	2.248	0.027	0.003	0.047
WMT	0.0272	0.041	0.667	0.506	-0.054	0.108
MCD	0.1435	0.057	2.536	0.013	0.031	0.256
USB	0.0164	0.051	0.322	0.748	-0.084	0.117
YUM	0.1469	0.047	3.114	0.002	0.053	0.240
NFLX	0.0972	0.021	4.539	0.000	0.055	0.140
JPM	0.1415	0.055	2.583	0.011	0.033	0.250
PFE	0.0546	0.033	1.662	0.099	-0.010	0.120
F	-0.0068	0.036	-0.187	0.852	-0.078	0.065
GM	-0.0105	0.027	-0.388	0.699	-0.064	0.043
COST	0.2176	0.059	3.713	0.000	0.101	0.334
JNJ	0.2414	0.056	4.350	0.000	0.131	0.351
Omnibus:	7.561	Durbin-Watson:	1.862			
Prob(Omnibus):	0.023	Jarque-Bera (JB):	8.445			
Skew:	0.400	Prob(JB):	0.0147			
Kurtosis:	3.978	Cond. No.	156.			

图 7. 多元 OLS 线性回归解

利用 `sklearn.linear_model.Ridge()` 函数，我们可以求解上述问题的岭回归参数。设定不同的 α 值，可以获得一系列岭回归参数。图 8 所示为随着 α 增大，岭回归参数变化。可以发现， α 增大时，参数逐步最大限度接近 0，但是不等于 0。这一点和本章后文将介绍的套索回归和弹性网络回归截然不同。

用残差平均值 MSE 来量化岭回归参数和 OLS 参数的差距：

$$\text{MSE}(\mathbf{b}_{\text{ridge}}, \mathbf{b}_{\text{OLS}}) = \frac{1}{D+1} \|\mathbf{b}_{\text{ridge}} - \mathbf{b}_{\text{OLS}}\|_2^2 \quad (6)$$

图 9 所示为随着 α 增大，岭回归参数和 OLS 参数的差距不断增大。

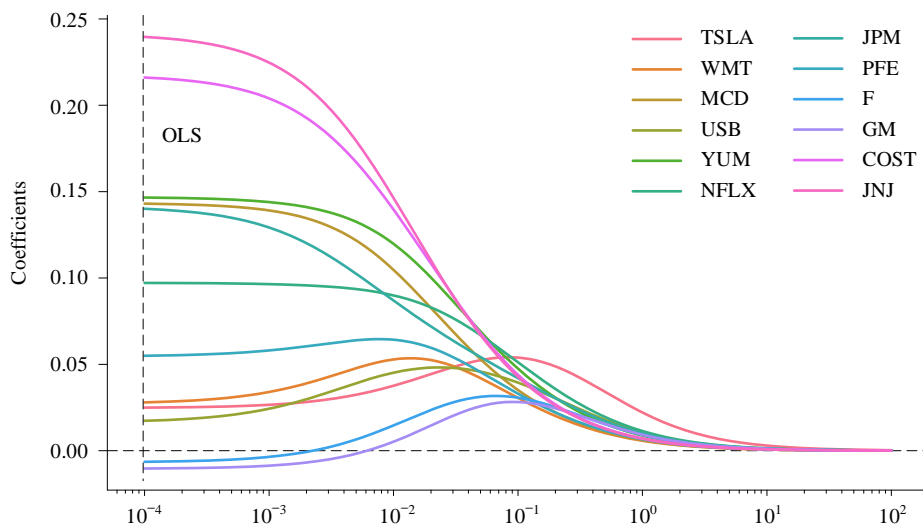
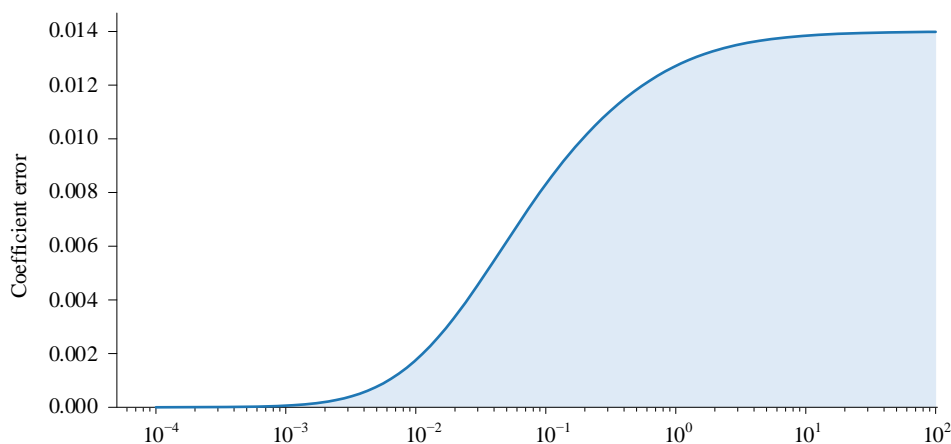
图 8. 随着 α 增大, 岭回归参数变化

图 9. 和 OLS 相比, 岭回归参数误差

Bk7_Ch05_01.ipynb 绘制本节前文图像。下面聊聊代码 1 中关键语句。

- a 创建了一个 `sklearn.linear_model.Ridge()` 的实例。
- b 创建了一个包含在对数尺度上均匀分布的 200 个 `alpha` 值的 NumPy 数组。
- c 中每次迭代, 设置 Ridge 回归模型的正则化参数为当前的 `alpha` 值。
- d 使用训练数据 `X_df` 和目标变量 `y_df` 进行拟合。
- e 将当前 `alpha` 值下的系数添加到列表 `coefs` 中。
- f 用 `sklearn.metrics.mean_squared_error()` 计算当前 `alpha` 值下的均方误差, 并添加到列表 `errors` 中。
- g 获取当前 `alpha` 值下的系数。
- h 创建一个 `DataFrame`, 其中包含了当前 `alpha` 值下的非截距项系数, 并设置相应的索引和列名。
- i 通过 `pandas.concat()` 合并 `DataFrame`。

本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com

```

from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error

a clf = Ridge()
  coefs = []
  errors = []
  coeff_df = pd.DataFrame()

b alphas = np.logspace(-4, 2, 200)

  for alpha_i in alphas:
c     clf.set_params(alpha=alpha_i)
d     clf.fit(X_df, y_df)
e     coefs.append(clf.coef_)
f     errors.append(mean_squared_error(clf.coef_,
                                      b.reshape(1, -1)))

g     b_i = clf.coef_
h     b_X_df = pd.DataFrame(data=b_i[:, 1:].T,
                           index = tickers[1:],
                           columns=[alpha_i])

i     coeff_df = pd.concat([coeff_df, b_X_df], axis = 1)

```

代码 1. α 对岭回归模型参数影响 | Bk7_Ch05_01.ipynb

多项式回归 + 岭正则

《编程不难》还介绍一个“多项式回归 + 岭回归”的例子。这个例子中，多项式回归次数较高会导致过拟合，而岭正则可以抑制过拟合。

图 10 所示为调整岭正则**惩罚因子** (penalty) α 对多项式回归模型的影响。显然，随着 α 不断增大，拟合得到的曲线变得更加“平滑”，这意味着模型变得更简单。表 1 给出在不同惩罚因子 α 条件下多项式模型解析式。

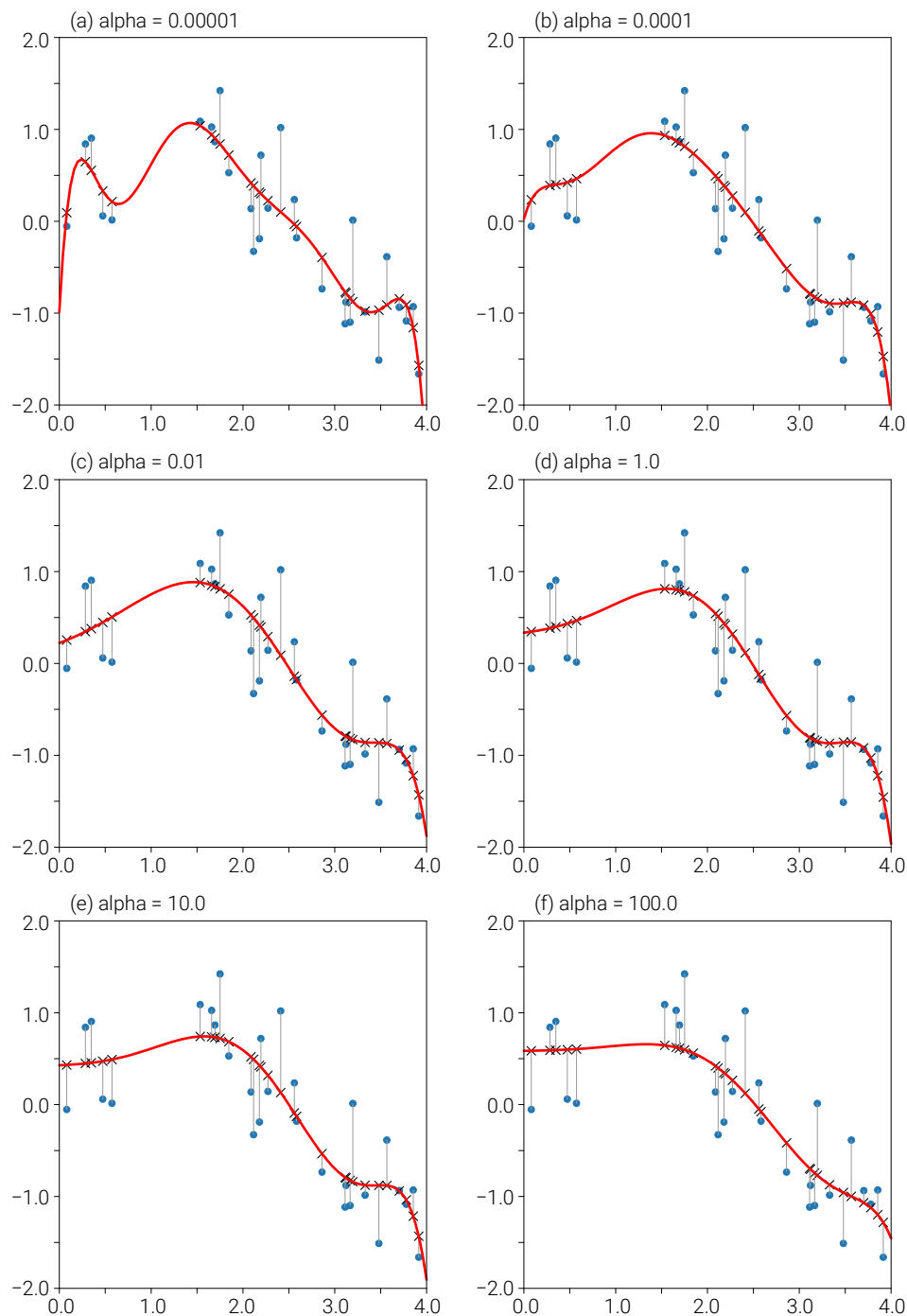
图 10. 岭正则化中惩罚因子 α 对多项式回归模型的影响，图片来自《编程不难》

表 1. 岭惩罚因子和多项式回归模型解析式，表格来自《编程不难》

惩罚因子 α	多项式回归模型
0.00001	$y = -0.985 + 18.400x^1 - 71.750x^2 + 122.612x^3 - 108.324x^4 + 53.620x^5 - 15.058x^6 + 2.243x^7 - 0.138x^8$
0.0001	$y = 0.026 + 3.491x^1 - 13.188x^2 + 24.668x^3 - 23.210x^4 + 12.008x^5 - 3.515x^6 + 0.547x^7 - 0.035x^8$
0.01	$y = 0.222 + 0.380x^1 + 0.149x^2 + 0.258x^3 - 0.391x^4 + 0.203x^5 - 0.093x^6 + 0.027x^7 - 0.003x^8$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

1.0	$y = 0.335 + 0.125x^1 + 0.132x^2 + 0.099x^3 + 0.019x^4 - 0.048x^5 - 0.033x^6 + 0.022x^7 - 0.003x^8$
10.0	$y = 0.428 + 0.045x^1 + 0.064x^2 + 0.070x^3 + 0.049x^4 - 0.008x^5 - 0.065x^6 + 0.030x^7 - 0.004x^8$
100.0	$y = 0.585 + 0.013x^1 + 0.020x^2 + 0.024x^3 + 0.019x^4 - 0.004x^5 - 0.029x^6 + 0.013x^7 - 0.002x^8$

5.3 几何角度看岭回归

从另外一个角度看岭回归，岭回归可以看做是 OLS 线性回归问题，加一个约束条件。

$$\begin{aligned} \arg \min_b \|y - Xb\|_2^2 \\ \text{subject to: } \|b\|_2^2 - c \leq 0 \end{aligned} \quad (7)$$

(7) 中的约束条件中 c 是一个阈值，就是把回归参数限制在一定范围之内，即：

$$b_0^2 + b_1^2 + b_2^2 + \dots + b_D^2 \leq c \quad (8)$$

注意，(7) 中阈值 c 越小，对应惩罚系数 α 越大。

不考虑常数系数， $D = 2$ 时，

$$b_1^2 + b_2^2 \leq c \quad (9)$$

上式为一个正圆面，圆心位于原点，半径为 \sqrt{c} 。OLS 对应的是旋转椭圆等高线和 (9) 正圆相切就是约束条件下优化解，也就是岭回归系数。

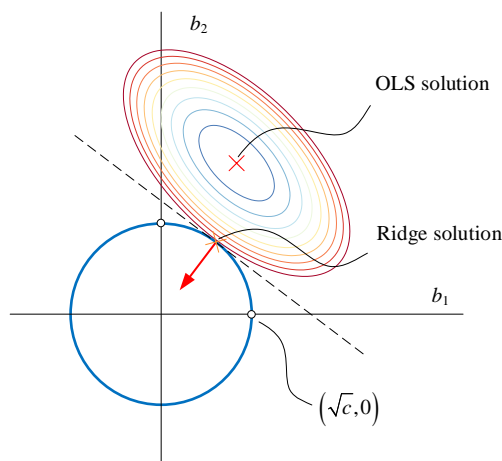
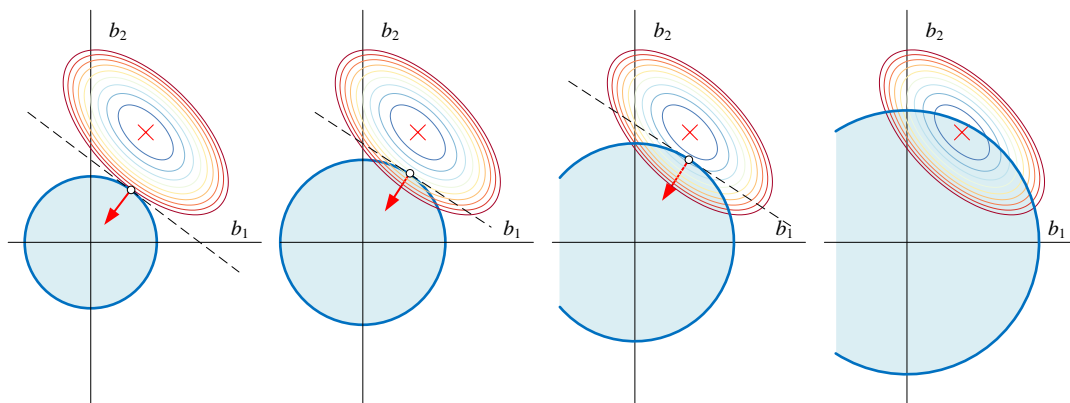


图 11. 约束角度看岭回归

图 12 所示为正圆面半径 \sqrt{c} 取不同值时，岭回归系数的优化解位置变化。

图 12. c 取不同值时，岭回归优化系数位置

多元 OLS 线性回归系数 \mathbf{b} 的解：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (10)$$

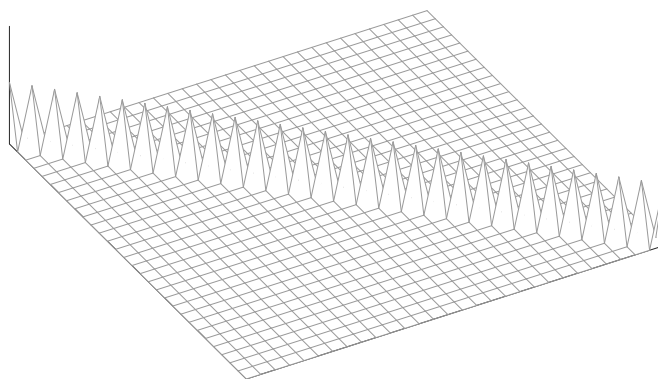
根据本书前文介绍的内容，OLS 线性回归的优化问题解存在且唯一的条件是 \mathbf{X} 列满秩。

如果，不满足 \mathbf{X} 列满秩这个条件，则表明 \mathbf{X} 列向量存在线性相关，即多重共线性。当 \mathbf{X} 列与列之间线性相关或者线性相关较大时， $\mathbf{X}^T \mathbf{X}$ 的行列式等于或接近于 0，无法求解(10) 中 $\mathbf{X}^T \mathbf{X}$ 一项的逆，会使得 OLS 解不稳定，

而岭回归线性回归系数 \mathbf{b} 的解为：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (11)$$

比较 (10)，可以发现 (11) 中变为求解 $\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I}$ 的逆；将 $\mathbf{X}^T \mathbf{X}$ 加上矩阵 $\alpha \mathbf{I}$ 变成非奇异矩阵并可以进行求逆运算。而 $\alpha \mathbf{I}$ 为对角矩阵，对角线上元素为 α ，其余为 0，形状酷似“山岭”，这也就是“岭回归”名称的由来。

图 13. $\alpha \mathbf{I}$ 对角矩阵引入的“山岭”

5.4 套索回归

斯坦福大学教授 Robert Tibshirani 在 1996 年首次提出将 L1 范数作为 OLS 正则项，得到 Lasso 模型。Lasso 是 least absolute shrinkage and selection operator 的缩写。

套索的优化目标函数为：

$$f(\mathbf{b}) = \underbrace{\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2}_{\text{OLS}} + \underbrace{\alpha \|\mathbf{b}\|_1}_{\text{L1 regularizer}} \quad (12)$$

图 14 所示为给定 α 条件下，(12) 如何构造得到套索回归目标函数参数曲面等高线图。如前文所述，(12) 目标函数中 OLS 部分对应椭圆抛物面，最小值点为红色 \times ；红色 \times 为二元 OLS 线性回归参数解的位置。(12) 中 L1 正则项曲面等高线对应旋转正方形，最小值点为蓝色 \times ，位于原点。

容易发现，黄色 \times 位于红色 \times 和蓝色 \times 之间；相对 OLS 线性回归参数红色 \times ，套索回归参数黄色 \times ，更靠近原点。

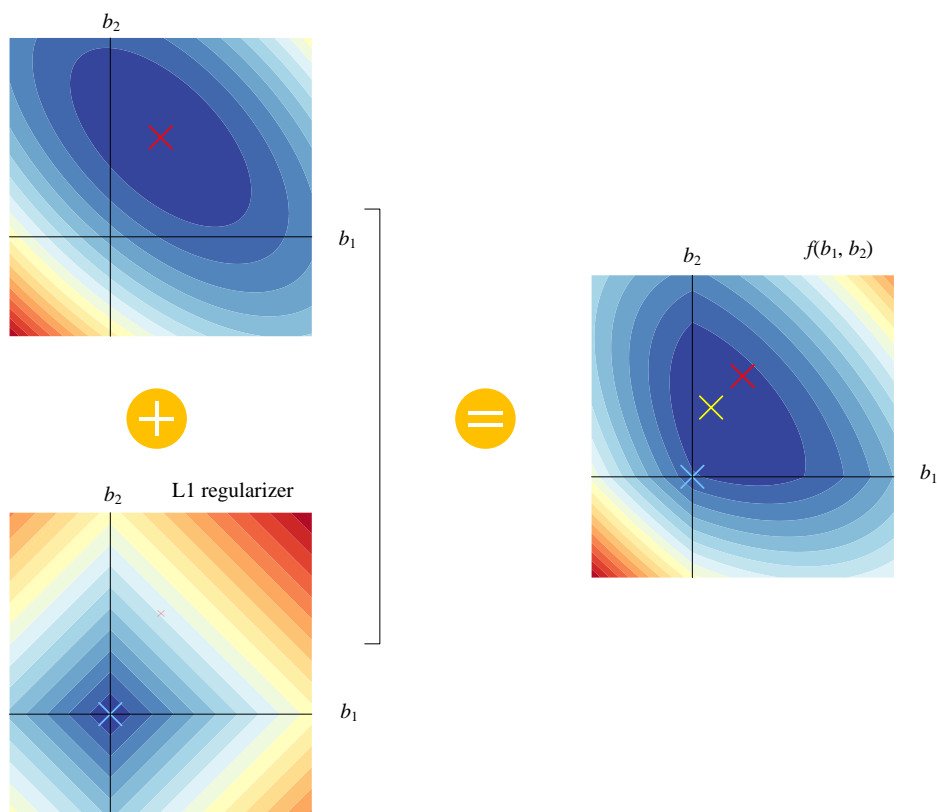
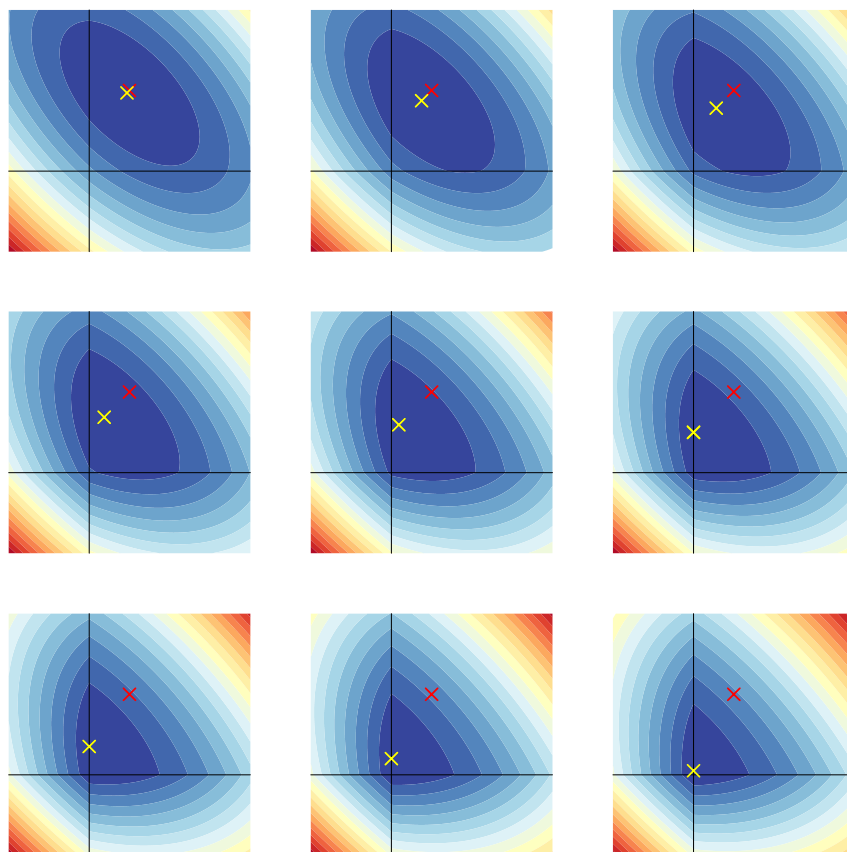


图 14. 构造套索回归优化问题参数曲面

图 15 所示为不断增大 α ，套索回归参数位置变化；可以发现套索回归采用 L1 正则化，可以导致参数估计结果为 0。

图 15. 不断增大 α ，套索回归参数位置变化

利用 `sklearn.linear_model.Lasso()` 可以获得套索回归的结果，利用本章前文的代码，将岭回归函数，换成套索回归函数，对于同一个问题，可以得到图 16。该图所示为随着 α 增大，套索回归参数变化。

观察图 16，可以发现在回归模型中， α 增大，一些特征快速收缩为 0，这个过程也是一个特征选择的过程。在套索回归中，系数越小表示对结果的影响越小，系数为 0 表示该特征没有对结果的影响，因此套索回归可以用于特征选择和降维。因此套索回归可以删除没有必要的特征，产生更为简洁的回归模型。特别地，`sklearn.linear_model.lars_path()` 函数可以用来生成套索回归参数轨迹图。

图 17 所示为和 OLS 相比，套索回归参数误差。

此外，请大家试着用套索回归完成图 10 这个多项式回归例子。

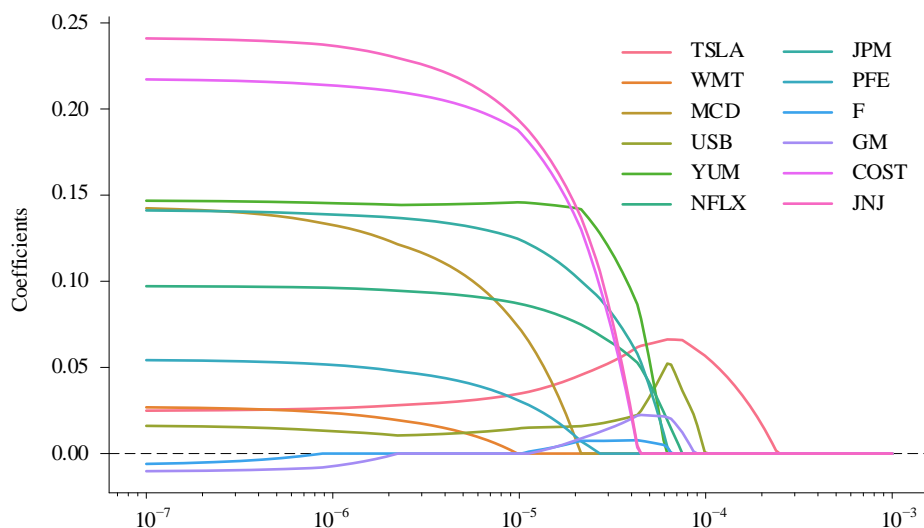
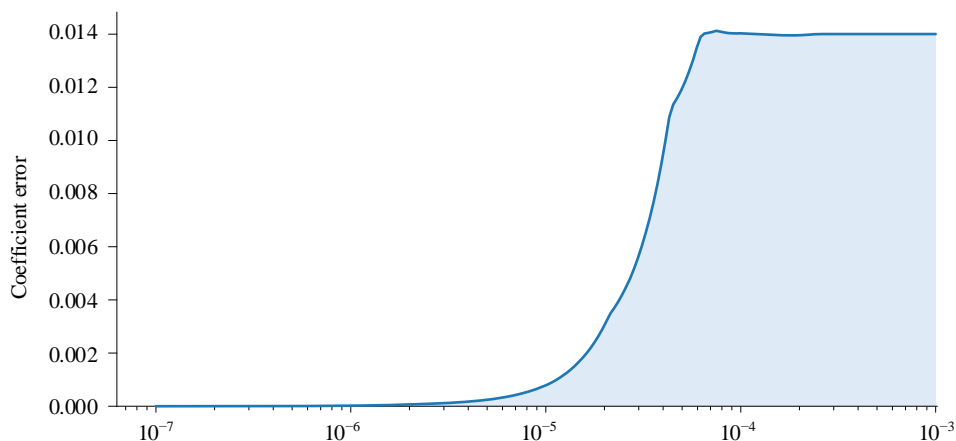
图 16. 随着 α 增大，套索回归参数变化

图 17. 和 OLS 相比，套索回归参数误差

5.5 几何角度看套索回归

类似地，本节从几何角度看套索回归。套索回归，可以看做是 OLS 线性回归问题，加一个约束条件：

$$\begin{aligned} \arg \min_b & \|y - Xb\|_2^2 \\ \text{subject to: } & \|b\|_1 - c \leq 0 \end{aligned} \quad (13)$$

(7) 中的约束条件中 c 也是一个阈值，即：

$$|b_0| + |b_1| + |b_2| + \dots + |b_D| \leq c \quad (14)$$

不考虑常数系数， $D = 2$ 时，

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$|b_1| + |b_2| \leq c \quad (15)$$

上式为一个旋转正方形，中心位于原点。OLS 对应的是旋转椭圆等高线可以和 (15) 旋转正方形相切，或在顶点处相交，如图 18 所示。

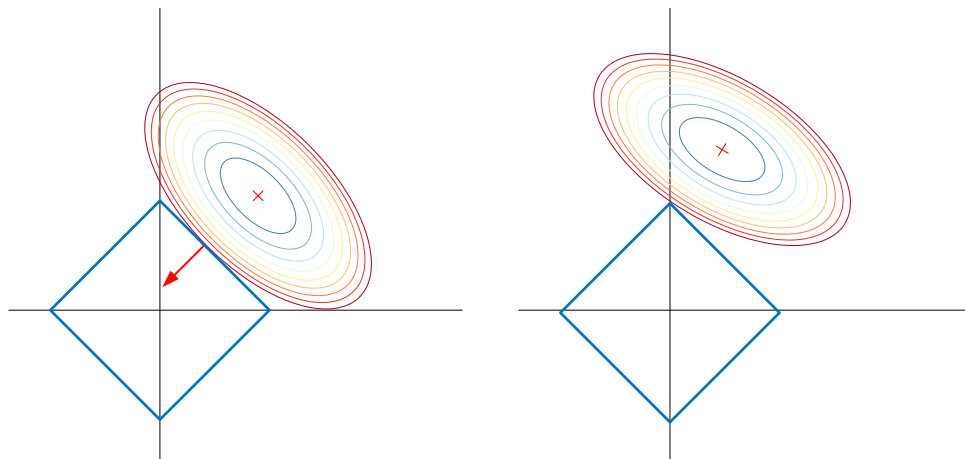


图 18. 套索回归的 L1 正则项

如图 19 所示，对于同一个 OLS 优化问题，不同的 c 阈值大小，会在不同位置得到套索回归系数解。前文说过，岭回归系数可以无限接近于 0，但是不等于 0；不同于岭回归，套索回归的参数可以直接为 0。套索回归参数的这种特点叫做**稀疏性** (sparsity)。稀疏性是指在套索回归中，某些特征系数被稀疏化为 0，使得模型参数更加简化和易于解释，同时也减少了数据维度，提高了模型的泛化能力。

当样本数据矩阵特征过多，但是只有少数特征对回归模型有贡献，去掉剩下的特征对模型没有什么影响。也就是说，回归模型只关注系数向量中非零项特征就足够了。因此，区别于岭回归，套索回归可以进行特征选择。

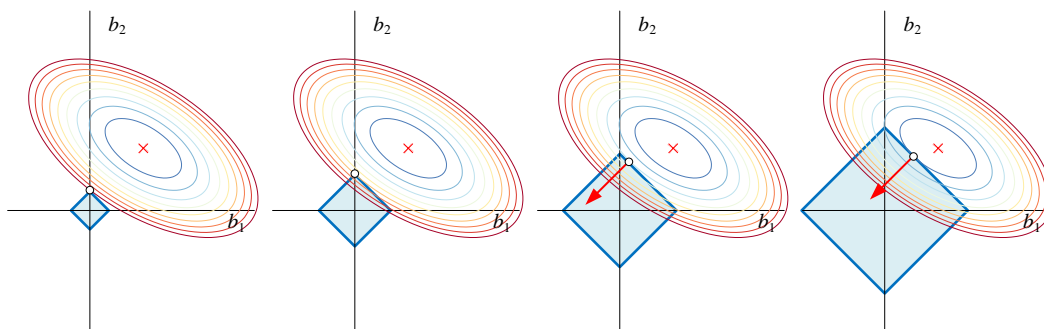


图 19. c 取不同值时，套索回归优化系数位置

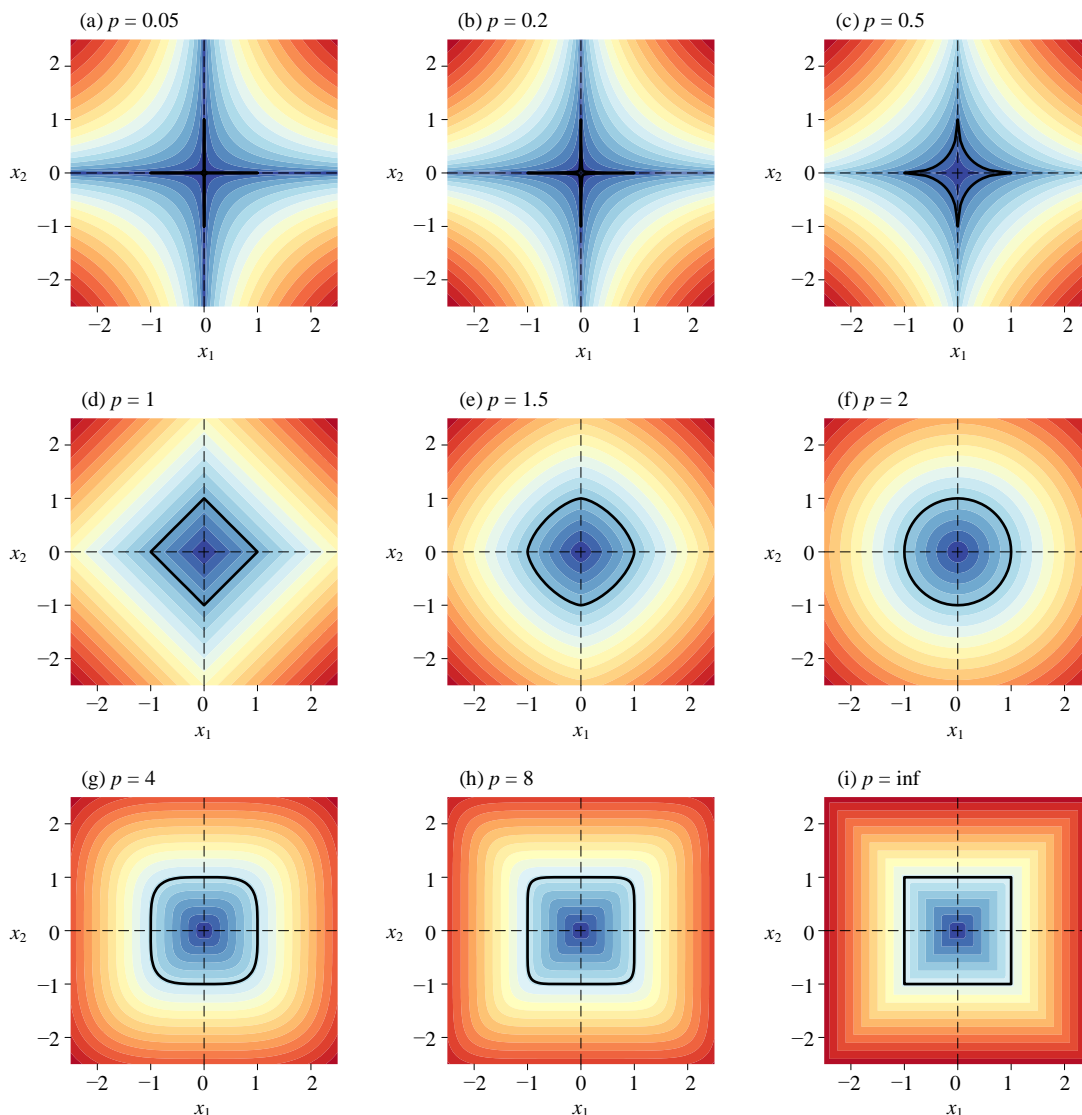


图 20. p 取不同值时, L^p 范数等高线形状变化; 注意, 严格来讲只有 $p \geq 1$ 才是范数

大家可能会问, 为什么 L^1 正则项会有这种稀疏性效果? 回顾丛书《矩阵力量》一书中给出的图 20。图 20 中给出, p 取不同值时, L^p 范数等高线形状变化。

可以发现, $p > 1$ 时, L^p 范数等高线形状连续光滑, 没有尖点。只有 $p \leq 1$ 时, 等高线图出现顶点尖点; 但是当 $p < 1$ 时, 目标函数为非凸函数, 优化问题求解困难。正是这个突出尖点的存在, 且满足凸优化问题, 让套索回归产生稀疏的向量解。

再次强调, 数学上严格来讲, 只有 $p \geq 1$ 才是 L^p 范数。

相信大家现在理解为什么, L^2 范数作为正则项, 无法产生稀疏性效果。二维平面下 L^2 正则项的等高线是正圆; 与正方形相比, 正圆根本没有棱角。因此 OLS 等高线和这个正圆相切时, 得到任意系数为 0 的机会几乎为零。这也就是为什么 L^2 正则化不具备稀疏性的原因。

以上结论不仅仅适用于二维, 三维甚至更多维度同样适用。图 21 比较三维空间的 L^1 和 L^2 正则项等高线曲面。



《数学要素》一本在超椭圆相关内容中介绍过图 21 图像。

本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com

图 21 (a) 中，L1 正则项存在大量突出尖点；这些尖点都对应着部分系数为 0。图 21 (b) 给出的正球体 (L2 正则)，任意一丁点扰动，比如计算误差、收敛等等，都会让回归系数不能恰好为 0。

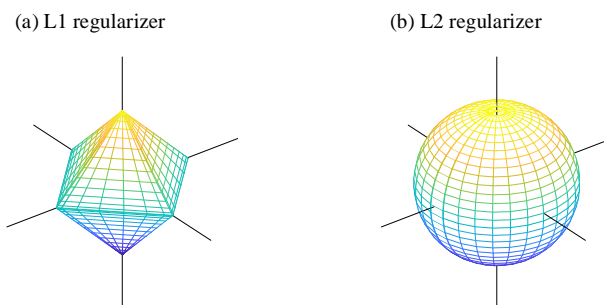


图 21. 三维空间的 L1 和 L2 正则项

此外，有些问题希望一些特征参数同时为 0，或者同时不为 0。这时可以设计，组 lasso (group lasso) 惩罚项来实现这一目标。与传统的 lasso 回归不同之处在于，组 lasso 回归在 L1 正则化项中增加了对特征分组的惩罚项。这个惩罚项是对组内系数的 L1 范数进行惩罚，从而鼓励组内特征系数共享相同的值或者趋近于零。因此，组 lasso 可以同时选择重要的特征和重要的特征组。这个方法在处理高维数据时特别有效，因为它可以减少特征的数量，避免过拟合，而且还可以保留组内特征之间的相关性。

图 22 所示为三维空间中两种 lasso 惩罚项结构。

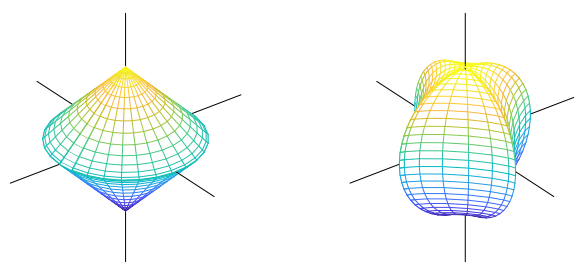


图 22. 三维空间中组 lasso 惩罚项

混合 L1 和 L2 正则项的弹性网络回归方法，可以克服 L2 正则项的不具备稀疏性这一缺点；这是我们下一节要介绍的内容。

5.6 弹性网络回归

弹性网络回归 (elastic net regression) 以不同比例同时引入 L1 和 L2 正则项，对应的目标函数为：

$$f(\mathbf{b}) = \underbrace{\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2}_{\text{OLS}} + \alpha \underbrace{\left(\rho \|\mathbf{b}\|_1 + \frac{(1-\rho)}{2} \|\mathbf{b}\|_2^2 \right)}_{\text{Elastic net regularizer}} \quad (16)$$

注意， α 为正则项惩罚系数，参数 ρ 用来调和 L1 和 L2 正则项的比例。

α 和 ρ 都是用户输入的数值。图 23 所示为构造弹性网络回归优化问题参数曲面等高线的过程。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

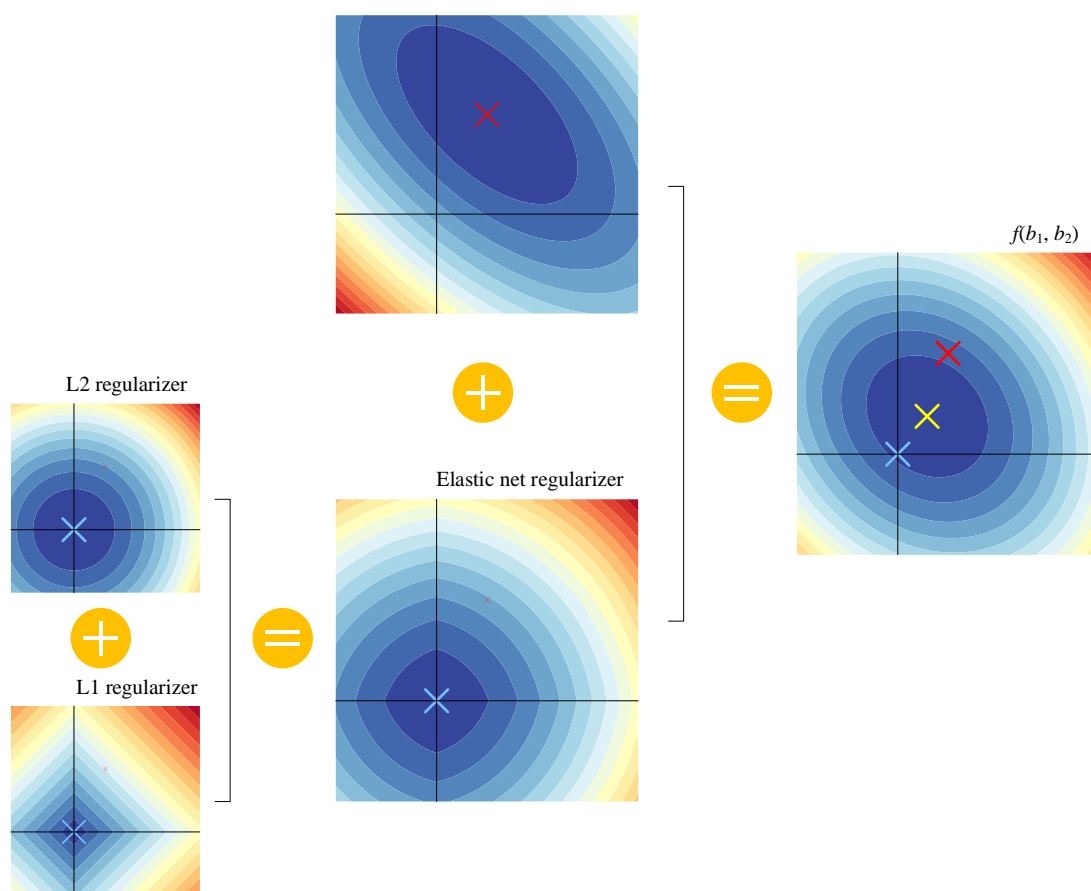


图 23. 构造弹性网络回归优化问题参数曲面等高线

图 24 所示为不断增大 α ，弹性网络回归参数位置变化。可以发现 α 增大，回归系数 b_1 不断靠近 0，甚至为 0。图 25 所示为回归系数运动轨迹，弹性网络回归系数靠近 0 的“速度”慢于套索回归。

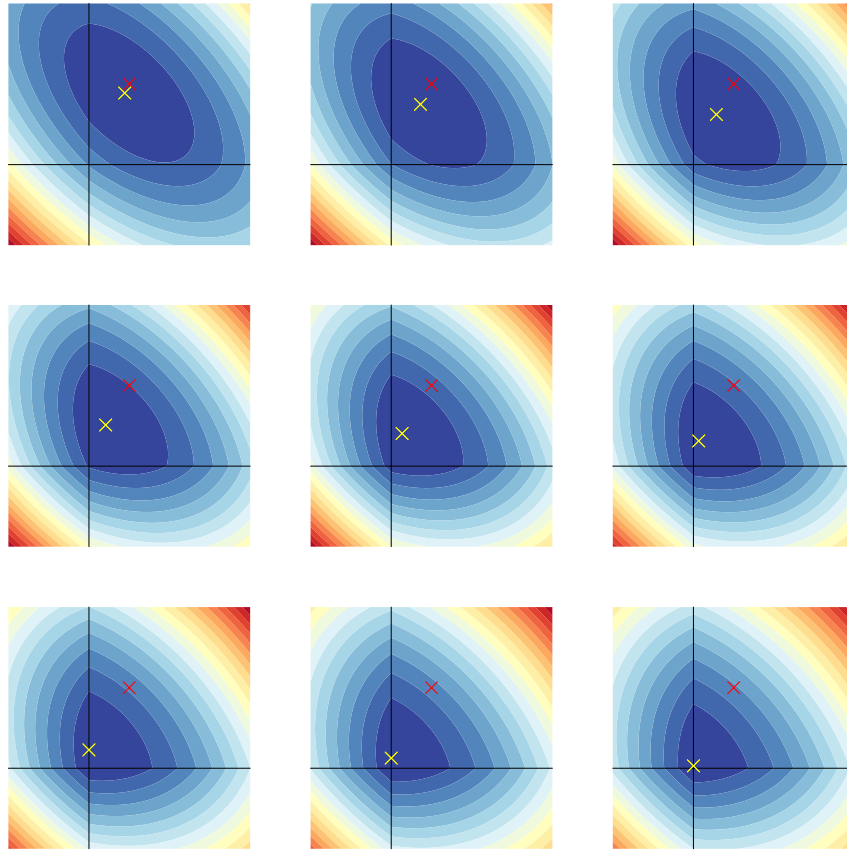


图 24. 不断增大 α , 弹性网络回归参数位置变化

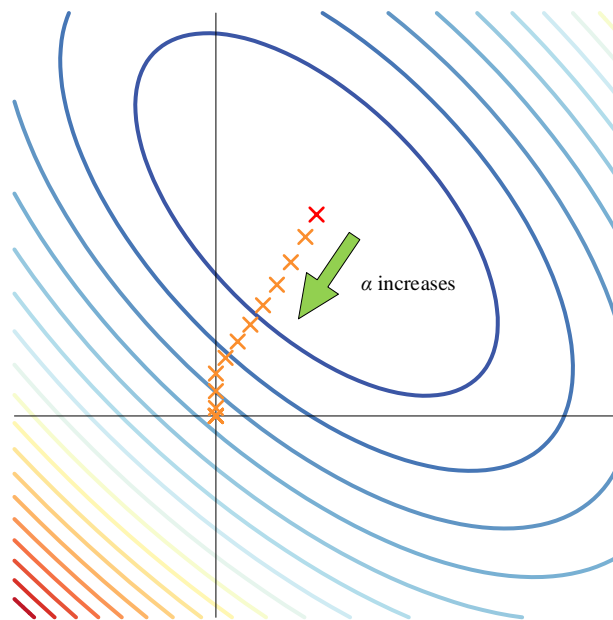


图 25. 不断增大 α , 弹性网络回归参数变化轨迹

本节前文介绍，参数 ρ 用来调和 L1 和 L2 正则项的比例；下面看一下参数 ρ 对弹性网络正则项形状的影响。图 26 和图 27 分别展示二维平面和三维空间中弹性网络正则项形状随 ρ 变化。 ρ 越大，弹性网络正则项越接近 L1，稀疏性越强； ρ 越小，弹性网络正则项越接近 L2，稀疏性越弱。

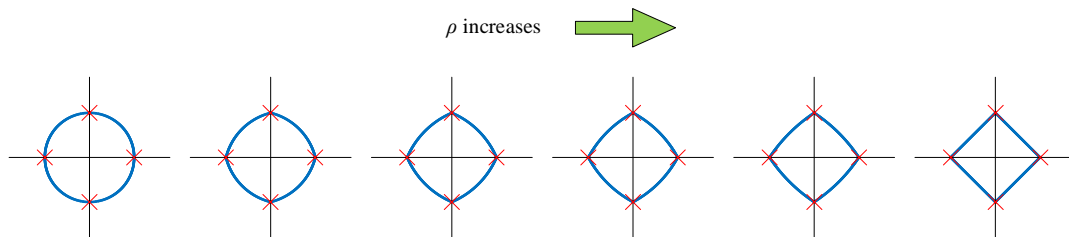


图 26. 不断增大 ρ ，二维平面弹性网络正则项等高线形状

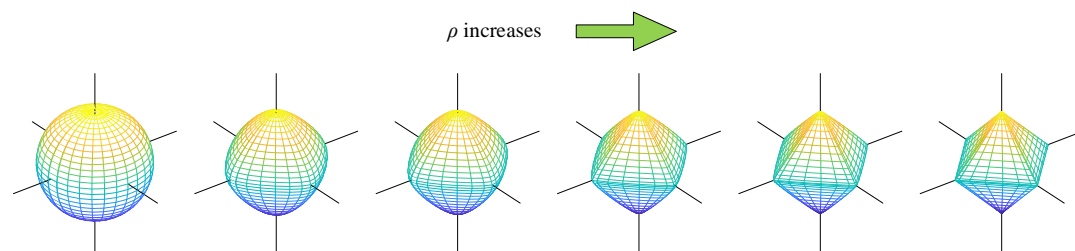


图 27. 不断增大 ρ ，三维空间弹性网络正则项等高面形状

图 28 所示为随着 α 增大，弹性网络回归参数变化，也就是弹性网络回归参数轨迹图。

注意，在这一过程中，参数 ρ 不变。

`sklearn.linear_model.ElasticNet()` 函数可以用来求解弹性网络回归问题。

此外，`sklearn.linear_model.enet_path()` 可以专门绘制弹性网络回归参数轨迹图。

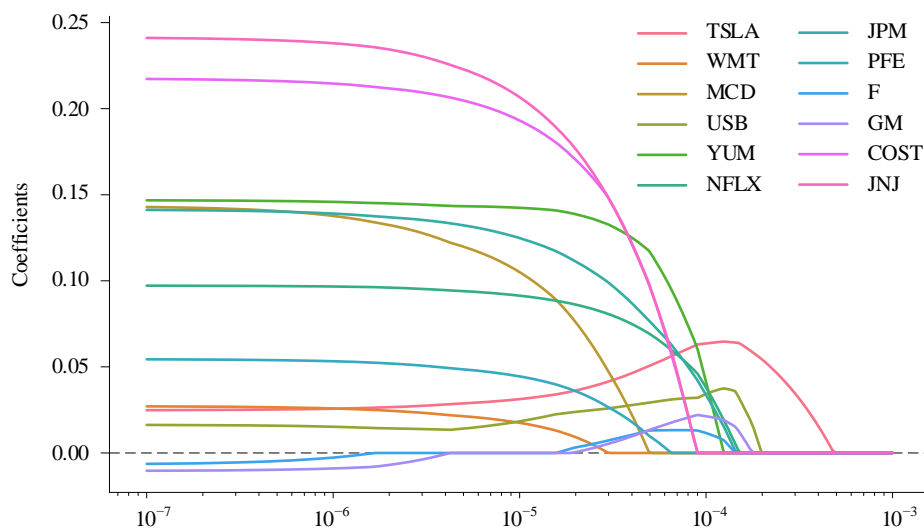


图 28. 随着 α 增大，弹性网络回归参数变化

图 29 比较套索回归和弹性网络回归参数随 α 变化；同样颜色的实线是套索回归参数，划线是弹性网络回归参数。容易发现，套索回归参数更快收缩到 0。弹性网络回归是套索回归和岭回归的结合体，它继承了套索回归的稀疏性，可以用来筛选特征，缩减无关参数。但是，由于引入岭回归 L2 正则项，弹性网络回归在淘汰特征的过程要慢于套索回归。

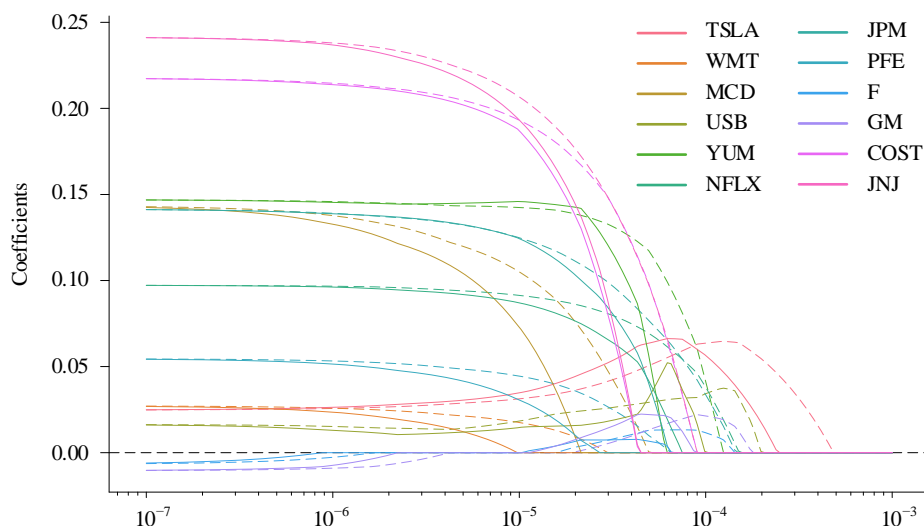


图 29. 比较套索回归和弹性网络回归参数随 α 变化

图 30 所示为和 OLS 相比，弹性网络回归参数误差。

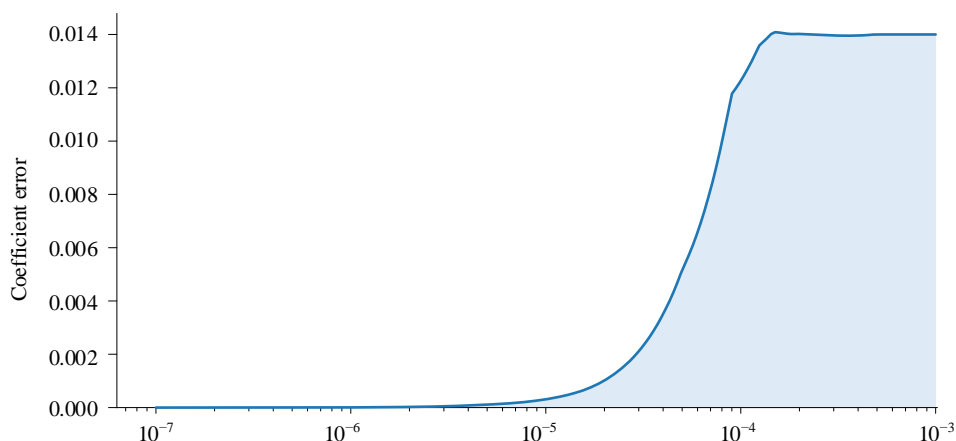


图 30. 和 OLS 相比，弹性网络回归参数误差

正则化是一种常用的机器学习技术，用于减小模型的复杂度和提高泛化能力。它通过在损失函数中添加一个正则项，强制模型参数的取值不要过大，从而避免模型过拟合。正则化技术包括 L1 正则化和 L2 正则化两种，L1 正则化将模型参数向 0 稀疏化，L2 正则化将模型参数平滑化，对于不同的数据集和

模型结构可以选择不同的正则化方法。正则化技术在实际应用中被广泛使用，可以提高模型的预测能力和稳定性，避免过拟合等问题。



推荐大家阅读 *Statistical Learning with Sparsity: The Lasso and Generalizations*。本书是稀疏统计学习专著。图书 PDF 文件可以免费从如下网址下载。

<https://web.stanford.edu/~hastie/StatLearnSparsity/>

有关岭回归，建议大家阅读 *Lecture notes on ridge regression*。下载地址如下：

<https://arxiv.org/abs/1509.09169>