第2章　Lasso

2025年7月

本讲关键问题

（1） 哪些应用场景会出现p/n较大的情况？

（2） 什么是Lasso, Ridge与弹性网？

（3） 为什么Lasso具有变量选择功能而Ridge没有？

（4） 如何选择Lasso中的惩罚系数？

（5） 用Lasso和Ridge进行预测需要满足什么条件？

关键词：稀疏性、Lasso、post Lasso、Ridge、弹性网、惩罚系数、交错鉴定法。

重点与难点

- 理解当p/n较大（高维解释变量）时，多元线性回归用于预测的方法与工作原理。这些方法包括Lasso，岭回归（ridge）与弹性网（elastic net）。
- 理解为何加入一定形式的惩罚项后，线性回归具有改善过度拟合与应对高维解释变量的能力，以及为何Lasso具有变量选择功能而Ridge没有此功能；
- 理解实际应用中如何正确使用Lasso，Ridge和弹性网等预测方法，以及选择恰当的惩罚系数。

一、解释变量个数p较大时的OLS

- 预测问题：

$$Y = m(X) + U,\ \ m(X) = X'\beta$$

当X的维数较大时，即使$m(X)$确实是X的线性函数，OLS的预测能力表现如何？

- 考虑以下数据生成过程：

p维度解释变量X服从联合正态分布$N(0, I_p)$；$U \sim^{p} N(0,1)$，$p \in \{10,20,30,40,50,60,70,80\}$，$Y = X'\beta + U$，$\beta_j = \frac{1}{j^{1.5}}, j = 1, \ldots, p$。

按照以上要求生成200个观测值。其中前100个观测值用于估计β，剩下100个观测值用于计算样本外MSE，样本内与样本外MSE的计算公式分别为：

$$\widehat{MSE} = \frac{1}{100}\sum_{I=1}^{100}(Y_i - X_i^{'}\hat{\beta})^2,$$

$$\widehat{MSFE} = \frac{1}{100}\sum_{I=101}^{200}(Y_i - X_i^{'}\hat{\beta})^2.$$

一、解释变量个数p较大时的OLS

以上过程重复500次，报告MSE和MSFE估计量的样本均值如下：

表2.1 $m(X)$是X的线性函数时OLS的预测能力

p	p/n	MSE	MSFE
10	0.1	0.8957	1.1180
20	0.2	0.7960	1.2403
30	0.3	0.6952	1.4277
40	0.4	0.5994	1.6743
50	0.5	0.4971	2.0084
60	0.6	0.4080	2.5147
70	0.7	0.3068	3.5387
80	0.8	0.2015	5.3231

一、解释变量个数p较大时的OLS

定理2.1.（OLS的预测性质）

设数据生成过程是$Y = X'\beta + U$，其中X是p维解释变量，$E(U|X) = 0$,$\hat{\beta}$是β的最小二乘估计量。则

$$E_X[(X'\hat{\beta} - X'\beta)^2] = (\hat{\beta} - \beta)'E(XX')(\hat{\beta} - \beta) \leq C \times \sigma^2 \times \frac{p}{n}$$

其中$E_X(\cdot)$表示对X求期望；上述不等式当$n \to \infty$时以概率1成立；常数C取决于观测样本(X, Y)的联合分布；$\sigma^2 = E(U^2)$。

定理2.1说明，即使Y的生成过程是X的线性函数，线性回归的预测性质好坏取决于p/n。 当p/n较大时，预测的准确度将下降。

定理2.1证明见：Belloni, Chetverikov, and Kato, 2015, Some new asymptotic theory for least squares series: pointwise and uniform, Chernozhukov results, Journal of Econometrics.

一、解释变量个数p较大时的OLS

➢ 什么情况下p/n会比较大？

（1）X的数量本来就很多。

- 例如为了估计一种商品A的需求函数（Y是A的销售量，X是产品价格，既包括A的价格，也包括同类产品的价格）。一种商品的同类产品数量可以是非常巨大的。（在线购物或者超市购物数据）。

（2）n的个数不大，此时即使X个数较少也会显得p/n很大。

- 例如第1章中研究过的growth.dta数据，其中N=90，但是解释变量超过60个。
- Imbens et al.（2001）收集了200多位购买彩票并且中奖的个人信息。作者要用这些信息研究彩票中奖（非劳动性收入）对劳动供给的影响。Imbens et al.（2001）一文中变量的描述性统计见下表。

表2.2 Imbens et al.（2001）搜集的彩票中奖者数据

Variables	Obs	Mean	Std.	Min	Max
Earnings 6 years after winning the lottery	202	11.465	14.339	0	44.816
Prize amount	237	57.369	64.842	1.139	484.790
age	237	46.945	13.797	23	85
gender (1 for male)	237	0.578	0.495	0	1
years of high school	237	3.603	1.071	0	4
years of college	237	1.367	1.601	0	5
winning year	237	6.059	1.294	4	8
number of tickers bought	237	4.570	3.282	0	10
work status after the winning (1 for working)	237	0.802	0.400	0	1
Earnings 1 year before winning the lottery	237	14.468	13.624	0	42.258
Earnings 2 years before winning the lottery	237	13.479	12.965	0	42.000
Earnings 3 years before winning the lottery	237	12.836	12.693	0	44.291
Earnings 4 years before winning the lottery	237	12.037	12.081	0	39.874
Earnings 5 years before winning the lottery	237	12.238	12.411	0	68.285
Earnings 6 years before winning the lottery	237	12.132	12.378	0	74.027

参考文献：

Imbens, G. W., Rubin, D. B., and Sacerdote, B. I., 2001, Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players, American Economic Review.

（3）X的个数原本不多，但是研究者还加入了X的平方项、高次项、交互项作为控制变量。

二、Lasso

- **如果真实的数据生成过程中X维数p很高，甚至$p > n$， 我们应当如何调整OLS算法适应这种情况？**

Lasso的基本思想：X或β的数量虽然很多，但是其中“只有一小部分”对Y有较大影响力；同时，研究者无须主观设定哪些变量“真正”具有影响力。这是“稀疏性”（sparsity）假设。

定义2.1（近似稀疏性 approximate sparsity）将$\beta = (\beta_1, \dots, \beta_p)$中各分量的绝对值从大到小排列后，记做$(\beta_{(1)}, \dots, \beta_{(p)})$。如果存在与无关的常数A和a, 使得$\beta_{(j)}$满足

$$|\beta_{(j)}| \leq A \times j^{-a}, a > \frac{1}{2}.$$

则称系数满足近似稀疏性。

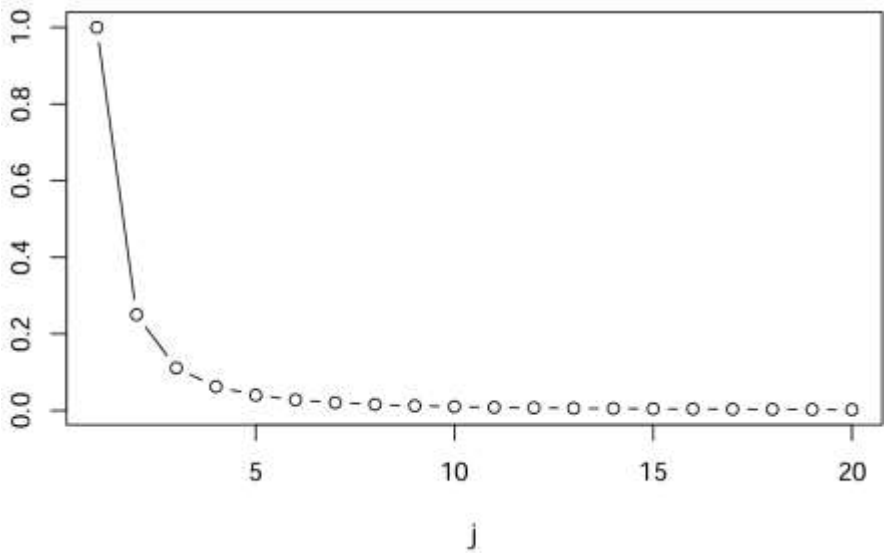

图2.1 系数的近似稀疏性

二、Lasso

Lasso （Least Absolute Shrinkage and Selection Operator）对系数向量施加“绝对值和”形式的惩罚：

$$\hat{\beta}_{lasso} = arg\min_{b} \sum_{i=1}^{n}(Y_i - X_i'b)^2 + \lambda \sum_{j=1}^{p} |b_j|\psi_j.$$

其中ψ_j表示各X_i之间的相对重要性。当每个X_i都被标准化为均值为0，标准差为1的变量之后，$\psi_j \equiv 1, j = 1, \ldots, p$。 一般情况下，

$$\hat{\psi}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2}.$$

$\lambda > 0$称为惩罚系数，它度量了对模型复杂度的限制。λ越大表示要求模型越简单（更多的β_j压缩至零）。

二、Lasso

数值模拟：p维解释变量X服从联合正态分布$N(0, I_p)$;$U \sim^d N(0,1)$，N=300,p=1000。数据生成过程为

$$Y = X'\beta + U, \beta_j = \frac{1}{j^{1.5}}, j = 1, \ldots, p.$$

表2.3 不同惩罚系数下Lasso的变量选择功能

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
	1	0.354	0.192	0.125	0.089	0.068	0.054	0.044
$\lambda^*/3$	0.843	0.284	0.076	0.141	0	0	0	0
$\lambda^*/2$	0.819	0.256	0.054	0.141	0	0	0	0
$\lambda^* = 0.17$ （由R程序给定）	0.750	0.173	0	0.093	0	0	0	0
$2\lambda^*$	0.562	0.004	0	0	0	0	0	0
$3\lambda^*$	0.381	0	0	0	0	0	0	0

二、Lasso

- 如何理解Lasso的工作原理？

不失一般性，以下仅考虑$\psi_j \equiv 1, j = 1, \ldots, p$的情形：$\hat{\beta}_{lasso} = arg\min_b \sum_{i=1}^n (Y_i - X_i'b)^2 + \lambda \sum_{j=1}^p [b_j]$.

估计量可以等价地表示成

$$\hat{\beta}_{lasso} = arg\min_b \sum_{i=1}^n (Y_i - X_i'b)^2, \text{ 满足} \sum_{j=1}^p [b_j] \leq \tau .$$

$$\hat{\beta}_{lasso} = arg\min_b \sum_{i=1}^n (Y_i - X_i'b)^2 + \lambda(\sum_{j=1}^p [b_j] - \tau).$$

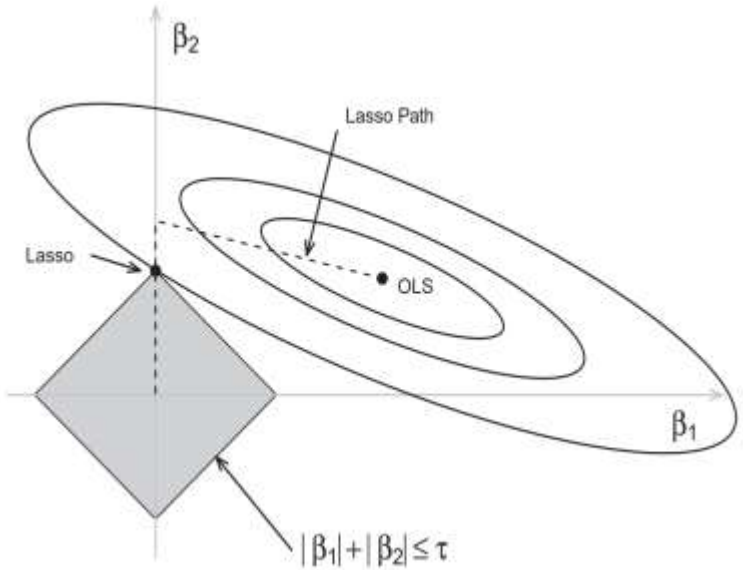

图2.2 p=2时Lasso选择变量的原理

二、Lasso

Lasso估计量的一阶条件：

$$-2X_n'(Y_n - X_n b) + \lambda \begin{bmatrix} \operatorname{sgn}(b_1) \\ \vdots \\ \operatorname{sgn}(b_p) \end{bmatrix} = 0.$$

为了说明Lasso的工作原理，假设矩阵X_n各列正交，即$X_n'X_n = nI_p$, 这相当于每个解释变量均值为零，方差为1，且相互独立。此时

$$\hat{\beta}_{ols} = X_n'Y_n/n,$$
$$\hat{\beta}_j = X_j'Y_n/n,$$
$$-2\left(\widehat{\beta_j} - \hat{\beta}_{L,j}\right) + \frac{\lambda}{n}\operatorname{sgn}\left(\hat{\beta}_{L,j}\right) = 0,$$
$$\hat{\beta}_{L,j} > 0, 2\left(\hat{\beta}_{L,j} - \widehat{\beta_j}\right) + \frac{\lambda}{n} = 0, \widehat{\beta_j} = \hat{\beta}_{L,j} + \frac{\lambda}{2n}.$$
$$\hat{\beta}_{L,j} < 0, 2\left(\hat{\beta}_{L,j} - \widehat{\beta_j}\right) - \frac{\lambda}{n} = 0, \widehat{\beta_j} = \hat{\beta}_{L,j} - \frac{\lambda}{2n}.$$

上式等价于

$$\widehat{\beta_j} > \frac{\lambda}{2n}, \hat{\beta}_{L,j} = \widehat{\beta_j} - \frac{\lambda}{2n}.$$
$$\widehat{\beta_j} < -\frac{\lambda}{2n}, \hat{\beta}_{L,j} = \widehat{\beta_j} + \frac{\lambda}{2n}.$$
$$-\frac{\lambda}{2n} < \hat{\beta}_j < \frac{\lambda}{2n}, \hat{\beta}_{L,j} = 0.$$

上述说明（1）λ越大，Lasso将更多系数估计量压缩至零；

（2）为了保持变量选择功能，λ应当随在n增大而增大。

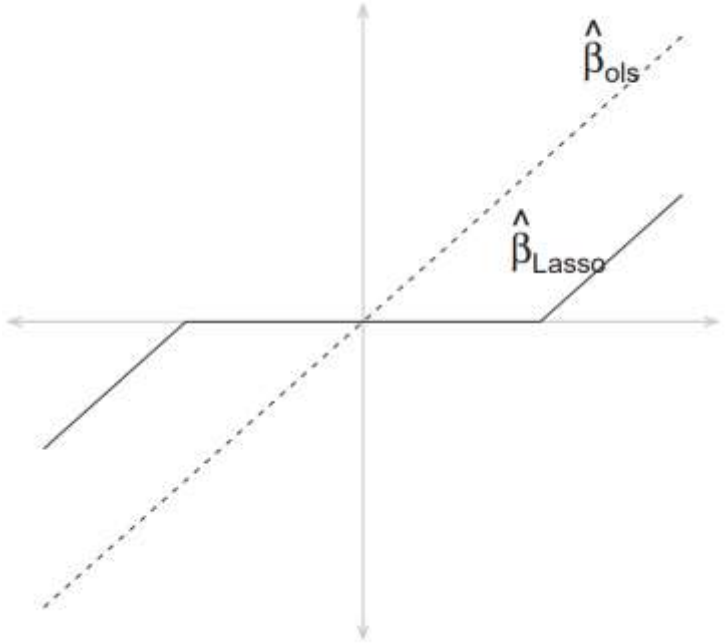

图2.3 Lasso的系数压缩功能

- Post Lasso

  第1步：用Lasso估计$Y = X'\beta + U$。记X_S是$\hat{\beta}_{lasso}$不等于零的解释变量。

  第2步：

$$\hat{\beta}_{plasso} = (\sum_{i=1}^{n} X_{si}X'_{Si})^{-1}(\sum_{i=1}^{n} X_{Si}Y_{Si}).$$

三、惩罚系数（模型复杂度）的选择

➢ 定理2.2（Lasso的预测性质）

设数据生成过程为$Y = X'\beta + U$, 其中X是p维解释变量，$E(U|X) = 0$，$\hat{\beta}_L$是β的Lasso估计量。惩罚系数λ满足

$$\lambda = 2c\hat{\sigma}\sqrt{n}z_{1-\frac{\gamma}{2p}}.$$

其中$c > 1$, $\hat{\sigma}^2$是$E(U^2)$的一致估计，γ是置信水平，一般选$\gamma = 0.05$, z_t是标准正态分布的第t分位数，即

$$P(N(0,1) < z_t) = t.$$

则

$$E_X\left[\left(X'\hat{\beta} - X'\beta\right)^2\right] \leq C \times \sigma^2 \times \frac{p}{n}.$$

其中$E_X(\cdot)$表示仅仅对X求期望；上述不等式当$n \to \infty$时以概率1成立；常数C取决于观测样本(X, Y)的联合分布；$\sigma^2 = E(U^2)$. s表示有效解释变量的维度；

$$s = s(a) = const \times A^{1/a} \times n^{1/2a}.$$

➢ 定理2.2给出了选择惩罚系数的方法之一：

$$\lambda = 2c\hat{\sigma}\sqrt{n}\Phi^{-1}(1 - \frac{\gamma}{2p}).$$

其中$\Phi^{-1}(\cdot)$是标准正态分布的分位数函数。一般选取$c = 1.1$, $a = 0.05$, σ^2通过迭代法得到。以上选择惩罚系数的方法详见：Belloni, Chen, Chernozhukov and Hansen, 2012, Sparse models and methods for optimal instruments with an application to eminent domain, Econometrica.

三、惩罚系数（模型复杂度）的选择

思考题：将定理2.2中

$$E_X\left[\left(X'\hat{\beta}-X'\beta\right)^2\right]\leq C\times\sigma^2\times\frac{s}{n}\times lnmax\{p,n\}$$

和定理2.1中OLS的预测性质

$$E_X\left[\left(X'\hat{\beta}-X'\beta\right)^2\right]\leq C\times\sigma^2\times\frac{p}{n}$$

进行比较。

- 除了根据（2.1）选择惩罚系数以外，也可以通过**交错鉴定法**选取。其步骤为：

（1）将全部数据分成K组（folds）。例如，K=10，每组中含观测值$\frac{n}{10}$。

（2）给定λ，每次留出K组的1组作为测试组。剩下的K-1组用于训练模型（Lasso）。并计算训练出的模型在测试集上的MSFE：

$$MSFE_k(\lambda)=\frac{1}{m_k}\sum_{i\in j_k}\left(Y_i-X_i\hat{\beta}_{-k}\right)^2$$

（3）重复上述步骤K次，轮流使每一块成为测试集。

（4）把上述K次中得到的均方预测误差的平均值作为模型预测精度的指标：

$$MSFE(\lambda)=\frac{1}{K}\sum_{k=1}^{K}MSFE_k(\lambda)$$

（5）求λ,使上述$MSFE(\lambda)$取得最小。此时的λ就是最佳惩罚系数。

四、其他带惩罚项的线性归

- 岭回归：

$$\hat{\beta}_{ridge}(\lambda) = \arg\min_{b} \sum_{i=1}^{n}(Y_i - X_i'b)^2 + \lambda \sum_{j=1}^{p}\left(b_j\right)^2.$$

$$\hat{\beta}_{ridge}(\lambda) = \left(\mathrm{X}_n'X_n + \lambda I_p\right)^{-1}(X_n'Y_n).$$

当$\lambda > 0$时，$\left(X_n'X_n + \lambda I_p\right)^{-1}$总是良好定义的。

$\hat{\beta}_{ridge}$的等价表示：受约束的OLS估计量$\hat{\beta}_{ridge}(\tau) = arg\ \min_{b'b<\tau} \sum_{i=1}^{n}(Y_i - X_i'b)^2.$

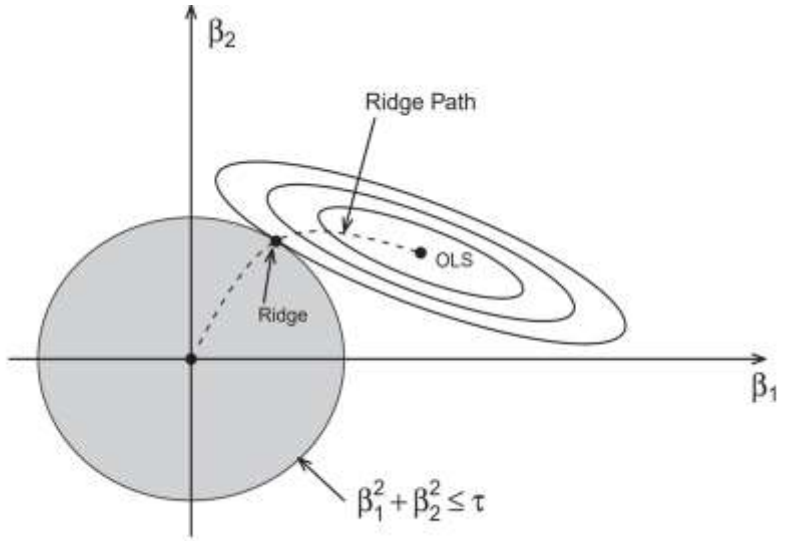

图2.4：岭回归工作原理

当$\lambda = 0$时， $\hat{\beta}_{ridge}$= $\hat{\beta}_{ols}$。当λ逐步增大，趋于无穷时，$\hat{\beta}_{ridge}$向原点靠拢（系数的绝对值被压缩）。

计算Lasso估计量最常用的程序包是R语言中的“glmnet”。

四、其他带惩罚项的线性归

- 为何岭回归没有变量选择功能？

在$X_n'X_n = nI_p$之下，

$$\hat{\beta}_{ridge}(\lambda) = \left(X_n'X_n + \lambda I_p\right)^{-1}(X_n'Y_n) = \left(nI_p + \lambda\right)^{-1} n\hat{\beta}_{ols} = \frac{\hat{\beta}_{ols}}{1+\frac{\lambda}{n}}.$$

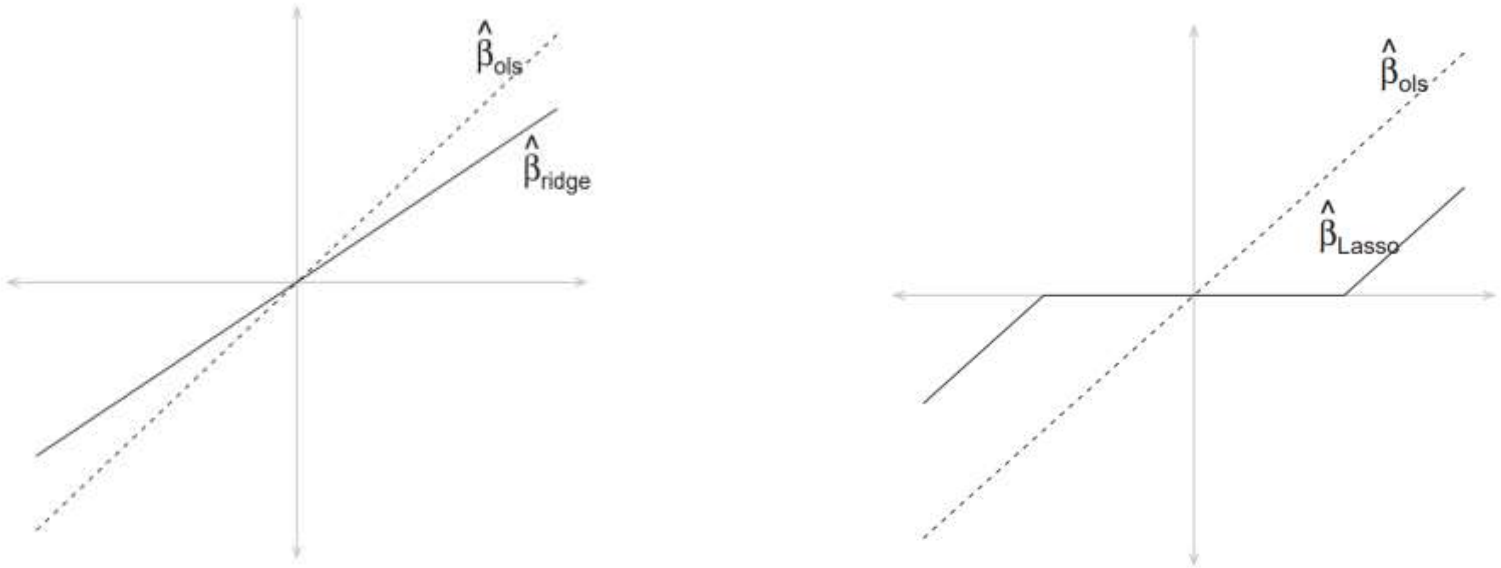

图2.5 Ridge和Lasso压缩系数方式对比

四、其他带惩罚项的线性归

- 弹性网（Elastic Net）估计量的目标函数介于Ridge和Lasso之间：

$$\hat{\beta}_{EN}(\lambda,\alpha) = arg\min_{b} \sum_{i=1}^{n}(Y_i - X_i'b)^2 + \lambda(\alpha\sum_{j=1}^{p} b_j^2 + (1-\alpha)\sum_{j=1}^{p}|b_j|)$$

$$\alpha \in [0,1]$$

弹性网在R中采用glmnet命令实现。

五、利用各种带惩罚项的线性回归预测经济增长

在第1章中，我们曾用以下六种线性模型预测各国经济增长：

模型编号	模型描述
1	包含变量：（1）人均实际GDP、（2）贸易开放度、（3）25岁及以上人口平均上学年限、（4）政治不稳定性、（5）劳动人口比重、（6）人口增长率、（7）政府实际消费支出与实际GDP比重
2	包含变量：模型1变量+（8）汇率、（9）65岁以上人口比重、（10）女性平均生育孩子数量、（11）25岁及以上男女平均上学年限之比
3	包含变量：模型1变量+这些变量的平方项
4	包含变量：模型2变量+这些变量的平方项
5	包含变量：模型3变量+这些变量的两两交互项
6	包含变量：模型4变量+这些变量的两两交互项

五、利用各种带惩罚项的线性回归预测经济增长

表1.4 模型（1）-（6）对经济增长的预测能力

设定	解释变量个数p	p/n	样本内MSE	K=3时MSFE估计量	K=5时MSFE估计量	K=10时MSFE估计量	K=90时MSFE估计量
1	8	4/45	0.00219	0.0033	0.00306	0.00297	0.00307
2	12	2/15	0.00207	0.00356	0.00326	0.00312	0.00320
3	15	1/6	0.00199	0.00357	0.00329	0.00320	0.00328
4	23	23/90	0.00170	0.00408	0.00361	0.00349	0.00358
5	36	2/5	0.00152	0.00486	0.00398	0.00375	0.00375
6	78	13/15	0.00024	0.00594	0.00517	0.00485	0.00505

五、利用各种带惩罚项的线性回归预测经济增长

由于Lasso, Ridge和弹性网都适用于解释变量个数大于样本容量的情形，因此我们可以用全部解释变量，以及他们的平方项，交互项来预测经济增长。

表2.4 模型（7）-（9）设定

模型编号	模型描述
7	所有解释变量的一次项
8	所有解释变量的一次项和平方项。
9	所有解释变量的一次项，平方项和两两之间交互项。

五、利用各种带惩罚项的线性回归预测经济增长

表2.5 采用Lasso, Ridge和弹性网，在模型（7）-（9）下预测经济增长

Lasso

设定	解释变量个数p	p/n	K=3时MSFE估计量	K=5时MSFE估计量	K=10时MSFE估计量	K=90时MSFE估计量
7	34	17/45	0.00260	0.00244	0.00244	0.00253
8	67	67/90	0.00260	0.00245	0.00250	0.00256
9	595	119/45	0.00260	0.00218	0.00232	0.00221

Ridge

设定	解释变量个数p	p/n	K=3时MSFE估计量	K=5时MSFE估计量	K=10时MSFE估计量	K=90时MSFE估计量
7	34	17/45	0.00257	0.00246	0.00250	0.00248
8	67	67/90	0.00260	0.00240	0.00250	0.00242
9	595	119/45	0.00244	0.00231	0.00247	0.00237

五、利用各种带惩罚项的线性回归预测经济增长

弹性网

设定	解释变量个数p	p/n	K=3时MSFE估计量	K=5时MSFE估计量	K=10时MSFE估计量	K=90时MSFE估计量
$\alpha = 0.2$						
7	34	17/45	0.00261	0.00246	0.00251	0.00256
8	67	67/90	0.00262	0.00252	0.00256	0.00254
9	595	119/45	0.00255	0.00222	0.00246	0.00234
$\alpha = 0.5$						
7	34	17/45	0.00260	0.00245	0.00246	0.00254
8	67	67/90	0.00260	0.00247	0.00252	0.00257
9	595	119/45	0.00260	0.00218	0.00238	0.00226
$\alpha = 0.8$						
7	34	17/45	0.00260	0.00244	0.00245	0.00253
8	67	67/90	0.00260	0.00245	0.00251	0.00256
9	595	119/45	0.00260	0.00218	0.00234	0.00222

拓展与参考文献

- Lasso的几何与计算性质最早由Tibshirani(1996)建立。Lasso的全称是“Least absolute shrinkage and selection operator”，也是由Tibshirani命名的。Lasso估计量存在着许多变型与衍生算法，本节只对其中最重要的类型进行了介绍。例如，与Lasso估计量密切相关的post Lasso估计量的性质由Belloni 和Chernozhukov（2013）建立。读者想要了解Lasso估计量的诸多拓展与变型，可以参考Hastie， Tibshirani 和Wainwright （2015）。
- 参考文献

Robert Tibshirani, 1996, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society, Series B.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright, 2015, Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall.

Alexandre Belloni and Victor Chernozhukov, 2013, Least squares after model selection in high-dimensional sparse models, Bernoulli.

Belloni, Chen, Chernozhukov and Hansen, 2012, Sparse models and methods for optimal instruments with an application to eminent domain, Econometrica.

Belloni, Chernozhukov, Chetverikov, and Kato, 2015, Some new asymptotic theory for least squares series: pointwise and uniform results, Journal of Econometrics.

谢 谢!