

变量选择案例分析

1. 摘要

本案例基于自适应 Lasso 进行改进，自适应 Lasso 是基于最小二乘估计，根据每个参数的大小进行不同程度的压缩。本案例调整自适应 Lasso 的参数参照，选入 Lasso 估计进行不同程度的压缩。

2. 改进措施

将自适应 Lasso 的参数参照改为 Lasso 估计，对 Lasso 估计进行不同程度的压缩，观察其变量选择的效果。

3. 数据来源

(1) 高维数据：来自于西南证券民企 100 的 1 月 22 日到 3 月 22 日的日线数据，共 100 个自变量，37 个样本数据。

(2) 低维数据：来自于西南证券上证 50 的 30min 数据，共 50 个自变量，320 个样本数据。

4. R 程序分析结果

(1) 低维数据剔除变量对比图：

```
> regsu.beta[which(regsu.beta==0)]
中泰证券..600918. 中国铁建..601186. 红塔证券..601236. 中国人保..601319.
0 0 0 0
中国石油..601857.
0
```

基于 BIC 准则剔除的变量

```
0.4751148 0.2083159 0.1756099
> step.beta[which(step.beta==0)]
中泰证券..600918. 中国铁建..601186. 中国人保..601319. 中国石油..601857.
0 0 0 0
```

基于逐步回归剔除的变量

```
> which(lassocp.beta==0)
中国人保..601319.
36
```

基于 Lasso 剔除的变量

```
[30] "民生银行..600016." "中国石化..600028." "三一重工..600031." "保利地产..600048." "中国联通..600050."
[6] "上汽集团..600104." "复星医药..600196." "恒瑞医药..600276." "万华化学..600309." "贵州茅台..600519."
[11] "恒生电子..600570." "海螺水泥..600585." "用友网络..600588." "海尔智家..600690." "闻泰科技..600745."
[16] "伊利股份..600887." "中泰证券..600918." "隆基股份..601012." "中国神华..601088." "工业富联..601138."
[21] "兴业银行..601166." "中国铁建..601186." "国泰君安..601211." "红塔证券..601236." "农业银行..601288."
[26] "中国平安..601318." "中国人保..601319." "新华保险..601336." "工商银行..601398." "中国建筑..601668."
[31] "华泰证券..601688." "京沪高铁..601816." "光大银行..601818." "中国中免..601888." "汇顶科技..603160."
[36] "药明康德..603259." "海天味业..603288." "韦尔股份..603501." NA
```

基于SCAD剔除的变量

```
> colnames(X)[which(en.beta==0)[-1]]
[1] "民生银行..600016." "中国石化..600028." "中信证券..600030." "三一重工..600031." "招商银行..600036."
[6] "保利地产..600048." "中国联通..600050." "上汽集团..600104." "复星医药..600196." "恒瑞医药..600276."
[11] "万华化学..600309." "恒生电子..600570." "用友网络..600588." "海尔智家..600690." "三安光电..600703."
[16] "闻泰科技..600745." "伊利股份..600887." "中泰证券..600918." "隆基股份..601012." "中信建投..601066."
[21] "中国神华..601088." "工业富联..601138." "兴业银行..601166." "中国铁建..601186." "国泰君安..601211."
[26] "红塔证券..601236." "农业银行..601288." "中国平安..601318." "新华保险..601336." "工商银行..601398."
[31] "中国太保..601601." "中国人寿..601628." "中国建筑..601668." "华泰证券..601688." "京沪高铁..601816."
[36] "光大银行..601818." "中国石油..601857." "中国中免..601888." "海天味业..603288."
> |
```

基于弹性网剔除变量

```
> which(beta.new==0)
上海机场..600009. 复星医药..600196. 贵州茅台..600519. 恒生电子..600570.
      3      12      15      17
闻泰科技..600745. 中泰证券..600918. 中信建投..601066. 红塔证券..601236.
      22      25      27      33
中国人保..601319. 新华保险..601336. 中国中免..601888. 汇顶科技..603160.
      36      37      46      47
海天味业..603288. 韦尔股份..603501. 兆易创新..603986.
      49      50      51
      ~~~~~
```

基于改进方法剔除变量

变量选择方法	残差结果
BIC准则	0.5512789
逐步回归	0.5173514
Lasso	0.5012446
SCAD	54.22772
弹性网	27.80082
改进方法	4.601461

低维数据残差对比图

(2) 高维数据结果对比:

```
> mean(abs(y.train-Z.train%*%scad.beta))
[1] 134.8675
> mean(abs(y.test-Z.test%*%scad.beta))
[1] 679.99
> |
```

高维数据下 SCAD 的残差结果

```
> colnames(X)[which(scad.beta!=0)[-1]]
[1] "X000963.华东医药" "X002032.苏.泊.尔" "X300033.同花顺"
[4] "X300122.智飞生物" "X300136.信维通信" "X300498.温氏股份"
[7] "X300759.康龙化成" "X601100.恒立液压"
> |
```

在 1 0 0 个自变量中 SCAD 保留的变量

```

> mean(abs(y.train-Z.train**en.beta))
[1] 0.9650685
> mean(abs(y.test-Z.test**en.beta))
[1] 133.7518
> colnames(X)[which(en.beta==0)[-1]]
character(0)
> colnames(X)[which(en.beta!=0)[-1]]

```

弹性网的残差结果

5. 结论与展望

(1) 从剔除的变量来看，改进的方法剔除的变量比 BIC、逐步回归和 Lasso 剔除得更多，说明压缩程度进一步加大，筛选出的剔除变量也均被 SCAD 和弹性网剔除，但 SCAD 和弹性网压缩强度更大，剔除的变量更多。

(2) 从残差结果来看，改进的方法剔除了更多的变量，残差自然也会更大一些，在减少了一定变量个数后，残差虽然比 BIC、逐步回归和 Lasso 更大，但是相比于 SCAD 和弹性网还是比较小。可能因为是低维数据，SCAD 和弹性网没有表现得很好。

(3) 在高维数据下，SCAD 剔除了较多变量而弹性网在本实验数据中没有剔除变量，残差结果来看，弹性网表现得更好。