

非负约束估计确定惩罚项权重系数的 LASSO 变量选择方法 在酿酒板块指数追踪中的应用

1、摘要：

根据指数的编制方法，行业板块指数是该行业板块下所有成分股股价的加权平均，权重与成分股的股本是有关的，也就是说某行业的板块指数是各个成分股的线性函数，可以通过建立行业板块指数与各个成分股的线性回归方程来估计成分股的权重，进行指数追踪。但基于同一行业板块的股票收益率之间强共线性的原因，传统线性回归构建的权重组合并不稳定；且在构建股票组合时没有全面地考虑到交易成本的影响。针对上述问题，本文考虑应用非负约束估计确定 LASSO 模型的惩罚项权重系数，以酿酒板块指数追踪为例，减少成分股个数的同时尽量降低对预测效果的影响。

2、使用方法与改进措施：

(1) LASSO 变量选择方法的基本思想：

简单来说，LASSO 回归就是在前面最小化目标函数 $Q(\beta)$ 的后面加了一个 L1 范数：

$$Q(\beta) = \|y - X\beta\|^2 + \lambda \|\beta\|_1$$

β 的求解即为：

$$\beta^* = \operatorname{argmin}_{\beta} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \sum_{k=1}^p |\beta_k|$$

用 Lasso 方法，可以把一些不重要变量的系数压缩为 0，既实现了较为准确的参数估计，也实现了变量选择（降维）。

(2) 自适应 LASSO 模型：

自适应 Lasso 模型是在 Lasso 算法的基础上改进而成，其系数估计由回归拟合的优良性和惩罚两部分构成。其中，惩罚项通过把较小的系数向 0 完全收缩达到剔除对应协变量的效果。

$$\beta^* = \operatorname{argmin}_{\beta} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \sum_{k=1}^p \hat{w}_k |\beta_k|$$

$$\hat{w}_k = |\tilde{\beta}_k|^{-\gamma}$$

其中 γ 为正常数， $\tilde{\beta}_k$ 为系数的初值估计量，不同情况下，初始值 $\tilde{\beta}_k$ 可以采用不同的估计方法。

自适应 Lasso 方法的一个突出优点是，通过调节权重系数，使得不同参数相应的惩罚不同，初始估计取倒数使得初始参数估计值较大的变量对应较小的权

重系数和较小的惩罚，初始估计值较小的变量对应较大的 权重系数和较大的惩罚。当某个自变量的初始估计值较大时，该自变量在自适应 Lasso 中权重系数就较小，即对应的惩罚就小，这样就保证了其以较高概率进入模型，即更容易被选择。

(3)改进措施:

因为成分股股价的权重应该是正数，先采用非负约束估计 $\tilde{\beta}_k$ ，确定 LASSO 惩罚项的权重系数 $\hat{w}_k = |\tilde{\beta}_k|^{-\gamma}$ ，用交叉验证方法确定 λ 的取值，训练带惩罚项系数的 LASSO 模型。

$$\begin{cases} \tilde{\beta} = \underset{\beta \geq 0}{\operatorname{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 \\ \hat{w}_k = |\tilde{\beta}_k|^{-1} \\ \beta^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \sum_{k=1}^p \hat{w}_k |\beta_k| \end{cases}$$

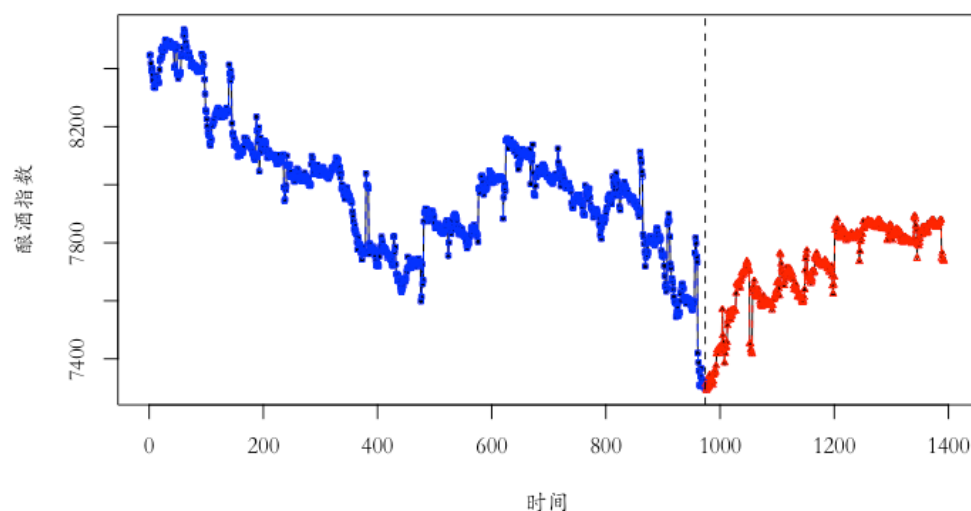
3、数据来源和采集时间:

西南证券金点子财富管理终端-行业板块: 2022. 1. 4-2. 18 酿酒板块及其成分股的 5 分钟 k 线的收盘价。共 1392 条数据，37 支成分股（协变量），1 个指数（因变量）。

采集于 2022 年 2 月 19 日。

4、R 程序分析结果:

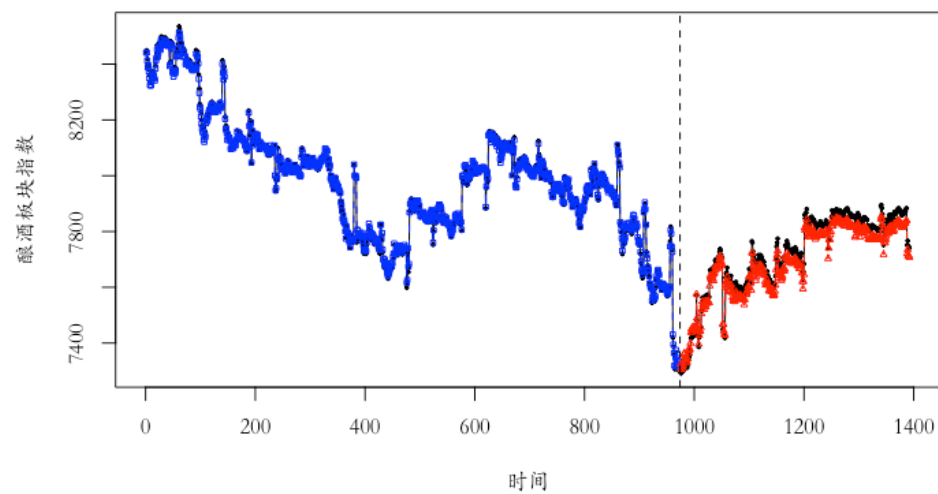
(1)Linear Regression:



MSE of training set: 0.4613258

MSE of testing set: 2.534829

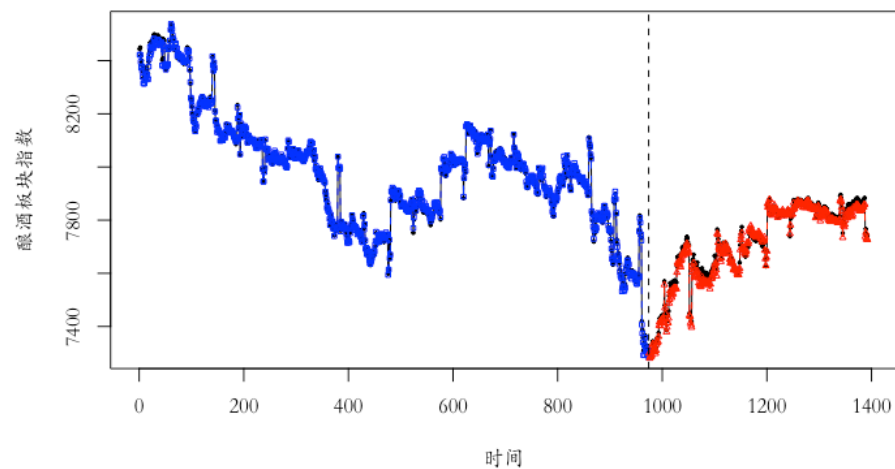
(2)LASSO:



MSE of training set: 5.765082

MSE of testing set: 32.01221

(3)Adaptive Lasso:



MSE of training set: 5.267726

MSE of testing set: 14.61813

5、结论和展望:

1) 结论:

线性回归模型在测试集上的预测效果最好, $MSE=2.53$, 但是包含 37 个协变量, 而且有很强的共线性问题; LASSO 进行变量选择后仅剩 10 个变量系数非零, 但测试集上的 $MSE=32.01$; 自适应的 LASSO 进行变量选择后仅剩 10 个变量系数非零, 且测试集上的 $MSE=14.61813$ 。

三个模型对比, 自适应的 LASSO 模型在起到降维作用的同时, 在预测效果上比原始的 LASSO 模型要好。

2)展望:

(1)选取数据的时间段可以更长一些,酿酒板块在春节前后的时期可能会波动较大,不是很具有代表性。

(2)可以对残差序列的异方差性和自相关性进行分析,还可以分析是否存在异常观测点导致回归系数的估计和检验不可靠;

(3)自适应的 LASSO 估计出的系数有负值,考虑加入 $\beta_k \geq 0$ 的约束会更好,使模型更具可解释性。

参考文献:

[1] 黄雅丹, 陆晓骏, 张大力. 具有模型一致性的自适应 Lasso 预测模型研究. 工业工程与管理.

<https://kns.cnki.net/kcms/detail/31.1738.T.20210611.1441.003.html>

[2]张怿瑾, 王成勇. 基于当前状态数据的加法风险模型的自适应 LASSO 回归选择 (英文)

[J]. 数学杂志, 2021, 41 (03):189-204. DOI:10.13548/j.sxzz.2021.03.001.

[3]秦晔玲, 朱建平. 基于自适应 Lasso 变量选择方法的指数跟踪[J]. 统计与决策, 2018, 34(16):141-145. DOI:10.13546/j.cnki.tjyjc.2018.16.035.

[4] Jianqing Fan, Yingying Fan, Emre Barut "Adaptive robust variable selection," The Annals of Statistics, Ann. Statist. 42(1), 324-351, (February 2014)