

基于中华120指数的股票投资组合策略

任 莎

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2023年4月3日; 录用日期: 2023年6月9日; 发布日期: 2023年6月16日

摘 要

指数追踪主要是对某一指数的表现进行复制或跟踪, 是一种流行的被动投资组合管理策略。随着中国股市的快速发展和不断成熟, 各项制度正在趋于规范和完善, 加之计算机技术的不断发展, 股指种类逐步丰富, 股指对于投资的作用显得更加重要, 很多投资者开始根据某些指数来进行股票的选择。本文以中华120指数及其成分股为研究对象, 采用R软件为计算工具, 运用逐步回归模型、Lasso、岭回归、两步估计法通过 C_p 准则、CV准则的方法得到了两个样本股空间, 最后在这两个样本股空间上对模型进行实证分析。数值实验表明基于Lasso的股票指数追踪模型具有较好的追踪误差及可解释性的优点。

关键词

指数追踪, 逐步回归, Lasso, 岭回归, 两步估计

Equity Portfolio Strategy Based on the China 120 Index

Sha Ren

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Apr. 3rd, 2023; accepted: Jun. 9th, 2023; published: Jun. 16th, 2023

Abstract

Index tracking is a popular passive portfolio management strategy that replicates or tracks the performance of a particular index. With the rapid development and continuous maturity of the Chinese stock market, various systems are becoming standardised and perfected; coupled with the continuous development of computer technology and the gradual enrichment of stock indices, the role of stock indices for investment has become more important and many investors have started to make stock selection based on certain indices. This paper takes the China 120 Index and its

constituent stocks as the research object, uses R software as the calculation tool, uses stepwise regression model, Lasso, ridge regression, two-step estimation method through the C_p criterion, CV criterion and other methods to obtain two sample stock spaces, and finally carries out empirical analysis of the model on these two sample stock spaces. Numerical experiments show that the Lasso-based stock index tracking model has the advantages of better tracking error and interpretability.

Keywords

Exponential Tracking, Stepwise Regression, Lasso, Ridge Regression, Two-Step Estimation

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

新中国的股票市场从上个世纪 80 年代就开始孕育, 当时, 国有企业的股份制改革可以认为是推动中国股市发展的重要因素。1990 年 12 月以后, 上海、深圳两个证券交易所相继成立, 至此以后的三十多个年头, 中国的股票市场便风雨兼程, 迅速发展壮大。

指数追踪是指用资本市场上若干个金融资产的组合来追踪市场上某一指数的表现。如用多只股票的组合通过权重的优化再配置来寻找部分股票构成的最优追踪组合。指数追踪主要是对某一指数的表现进行复制或跟踪, 指数复制的 3 种方法分别是完全复制法、抽样复制法和优化选样法。目前, 针对具有高精度、低交易成本, 且能保证追踪组合的高流动性的指数追踪技术的相关研究具有重要的意义[1]。

国内外学者针对此问题做了大量研究, 如李俭富等学者在 2005 年介绍了指数跟踪技术中跟踪组合的复制方法、研究的内容和指标, 对指数跟踪的基本模型、优化算法和其理论发展进行评述, 最后指出指数跟踪技术研究的发展方向[2]。进一步在 2006 年从系统风险控制的角度构造优化跟踪组合, 比较了分层抽样复制法、完全复制法和优化复制法构造的指数基金的实际跟踪效果[3]。Beasley (2008)提出了用于索引跟踪以及增强索引的混合整数线性规划公式, 将指数追踪技术的测试数据库从之前的 6 个数据集进一步扩充到 8 个数据集, 同时还考虑了费用对追踪效果的影响[4]。还有些学者提出使用主成分分析方法(PCA)来选取股票, 根据报表数据利用 R 软件使用 PCA 进行分析。同时利用 PCA 得出了一种综合排名, 结果表明, 评估比较合理, 能够反映上市公司的综合水平[5]。但是传统的 PCA 通常不具有稀疏性, 从而不能满足实际应用的需要。

随着指数型基金的迅速发展, 对指数追踪的研究也逐渐深入。陈春锋等学者在 2004 年以上证 180 指数进行实证分析, 对如何构建指数复制的模型进行研究, 并对不同复制方法的复制效果进行了研究[6]。在此基础上, 陈伟忠等学者在 2005 年从组合构建和调整等各个方面入手, 对指数优化复制中的流动性问题进行改进, 并研究了流动性问题对指数追踪的影响程度[7]。以及顾明庆等学者在 2023 年通过建立数学模型, 运用 LASSO 算法对成分股进行筛选, 进而构建能够复制该指数的投资组合[8]。

本文在解决股票指数追踪问题时, 采用不同的变量选择方法和指数追踪方法, 从而进行对比, 选择最优模型进行指数追踪。本文将以中华 120 指数为研究对象, 对传统的指数化投资方法进行实证研究。

2. 研究方法

2.1. 逐步回归

逐步回归是一种线性回归模型自变量选择方法，其基本思想是将变量一个一个引入，引入的条件是其偏回归平方和经验是显著的。同时，每引入一个新变量后，对已入选回归模型的老变量逐个进行检验，将经检验认为不显著的变量删除，以保证所得自变量子集中每一个变量都是显著的。此过程经过若干步直到不能再引入新变量为止。这时回归模型中所有变量对因变量都是显著的。

逐步回归法选择变量的过程包含两个基本步骤：一是从回归模型中剔出经检验不显著的变量，二是引入新变量到回归模型中，常用的逐步型选元法有向前法和向后法。向前法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。具体步骤如下：

步骤 1：对 p 个回归自变量 $X_1, X_2, \dots, X_p$ ，分别同因变量 Y 建立一元回归模型

$$Y = \beta_0 + \beta_i X_i + \varepsilon, i = 1, \dots, p \quad (1)$$

计算变量 X_i ，相应的回归系数的 F 检验统计量的值记为 $F_1^{(1)}, \dots, F_p^{(1)}$ ，取其中的最大值 $F_{i_1}^{(1)}$ ，即

$$F_{i_1}^{(1)} = \max \{F_1^{(1)}, \dots, F_p^{(1)}\} \quad (2)$$

对给定的显著性水平 α ，记相应的临界值为 $F^{(1)}$ ， $F_{i_1}^{(1)} \geq F^{(1)}$ ，则将 X_{i_1} 引入回归模型，记 I_1 为选入变量指标集合。

步骤 2：建立因变量 Y 与自变量子集 $\{X_{i_1}, X_1\}, \dots, \{X_{i_1}, X_{i-1}\}, \{X_{i_1}, X_{i+1}\}, \dots, \{X_{i_1}, X_p\}$ 的二元回归模型，共有 $p - 1$ 个。计算变量的回归系数 F 检验的统计量值记为 $F_k^{(2)} (k \notin I_1)$ ，选其中最大者，记为 $F_{i_2}^{(2)}$ ，对应自变量脚标记为 i_2 ，即

$$F_{i_2}^{(2)} = \max \{F_1^{(2)}, \dots, F_{i-1}^{(2)}, F_{i+1}^{(2)}, \dots, F_p^{(2)}\} \quad (3)$$

对给定的显著性水平 α ，记相应的临界值为 $F^{(2)}$ ， $F_{i_2}^{(2)} \geq F^{(2)}$ ，则变量 X_{i_2} 引入回归模型。否则，终止变量引入过程。

步骤 3：考虑因变量对变量子集 $(X_{i_1}, X_{i_2}, X_{i_k})$ 的回归重复步骤 2。

依此方法重复进行，每次从未引入回归模型的自变量中选取一个，直到经检验没有变量引入为止[9]。

后退法与前进法相反，开始时先拟合包含所有自变量的回归方程，并预先指定留在回归方程中而不被剔除的自变量的假设检验标准。然后按自变量对应变量 Y 的贡献大小从小到大进行检验，对无统计学意义的自变量依次剔除。每剔除一个自变量，都要重新计算并检验尚未被剔除自变量对应变量 Y 的贡献并决定是否剔除对模型贡献最小的自变量。重复上述过程，直到回归方程中的自变量均符合留在方程中的给定标准，没有自变量可被剔除为止。在整个过程中只考虑剔除自变量，自变量一旦被剔除，则不再考虑引入回归方程[10]。

2.2. Lasso

Lasso 是由 1996 年 Robert Tibshirani 首次提出，全称 Least absolute shrinkage and selection operator。该方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。Lasso 的基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而能够产生某些严格等于 0 的回归系数，得到可以解释的模型，其数学表达式如下：

$$B_{LASSO} = \arg_B \min \left\{ \left\| Y - \sum_{j=1}^p X_j B_j \right\| \right. \quad (4)$$

$$\left. s.t. \sum_{j=1}^p |B_j| \leq t \right.$$

其中 $t > 0$, 是调整参数, 通过控制调整参数 t 可以实现对总体回归系数的压缩。 t 值的确定可以利用 Efron 和 Tibshirani (1993) 提出的交叉验证法来估计。这个数学表达式还等价于最小化下述惩罚最小二乘法:

$$B_{LASSO} = \arg_B \min \left\{ \left\| Y - \sum_{j=1}^p X_j B_j \right\|^2 + a \sum_{j=1}^p |B_j| \right\} \quad (5)$$

其中 a 与 t 一一对应, 可以互相转换。Lasso 方法的主要优势在于其对参数估计较大的变量压缩较小, 而参数估计较小的变量压缩成 0, 并且 Lasso 分析的参数估计具有连续性, 适用于高维数据的模型选择。Tibshirani 在 2005 年提出了 Fused Lasso 方法, 这个估计方法满足了模型系数以及系数差分的稀疏性, 使得邻近系数间更加平滑。

2.3. 岭回归

岭回归(ridge regression, Tikhonov regularization)是一种专用于共线性数据分析的有偏估计回归方法, 实质上是一种改良的最小二乘估计法, 通过放弃最小二乘法的无偏性, 以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法, 对病态数据的拟合要强于最小二乘法。对于有些矩阵, 矩阵中某个元素的一个很小的变动, 会引起最后计算结果误差很大, 这种矩阵称为“病态矩阵”。有些时候不正确的计算方法也会使一个正常的矩阵在运算中表现出病态。对于高斯消去法来说, 如果主元上的元素很小, 在计算时就会表现出病态的特征。

回归分析中常用的最小二乘法是一种无偏估计。对于一个适定问题, X 通常是列满秩的

$$X\theta = y \quad (6)$$

采用最小二乘法, 定义损失函数为残差的平方, 最小化损失函数

$$\|X\theta - y\|^2 \quad (7)$$

上述优化问题可以采用梯度下降法进行求解, 也可以采用如下公式进行直接求解

$$\theta = (X^T X)^{-1} X^T y \quad (8)$$

当 X 不是列满秩时, 或者某些列之间的线性相关性比较大时, $X^T X$ 的行列式接近于 0, 即 $X^T X$ 接近于奇异, 上述问题变为一个不适定问题, 此时, 计算 $(X^T X)^{-1}$ 时误差会很大, 传统的最小二乘法缺乏稳定性与可靠性。

为了解决上述问题, 我们需要将不适定问题转化为适定问题: 我们为上述损失函数加上一个正则化项, 变为

$$\|X\theta - y\|^2 + \|\Gamma\theta\|^2 \quad (9)$$

其中, 我们定义 $\Gamma = \alpha I$, 于是:

$$\Gamma = \alpha I \quad (10)$$

随着 α 的增大, $\theta(\alpha)$ 各元素 $\theta(\alpha)_i$ 的绝对值均趋于不断变小, 它们相对于正确值 θ_i 的偏差也越来越大。 α 趋于无穷大时, $\theta(\alpha)$ 趋于 0。其中, $\theta(\alpha)$ 随 α 的改变而变化的轨迹, 就称为岭迹。实际计算中

可选非常多的 α 值，做出一个岭迹图，看看这个图在取哪个值的时候变稳定了，那就确定 α 值了。

岭回归是对最小二乘回归的一种补充，它损失了无偏性，来换取高的数值稳定性，从而得到较高的计算精度。

3. 数据来源

3.1. 中华 120 指数

本文数据来源 Choice 金融终端所下载的 2022 年 4 月 11 日~2023 年 4 月 7 日的中华 120 指数，除去部分时间数据缺失，共计 243 天数据。中华交易服务中国 120 指数(简称中华 120)，是涵盖中港两地大市值股份的跨境指数，由中华交易服务委托中证指数有限公司进行编制，并负责指数日常管理及指数发布服务。以人民币为实时报价，每日收市提供港元报价。

如何有效的利用股指期货的做空功能实现风险的对冲是眼下一个很实际的研究课题。作为基金管理者，要全部投资所有的 120 只成份股虽然可行，但是只选择部分成份股进行投资，如果要保证持有的股票组合与股指期货空单形成完全对冲，选择的股票组合必须完全吻合中华 120 指数的走势。

中华 120 成份股由 80 家流动性最高、市值最大的沪深交易所上市 A 股以及 40 家流动性最高、市值最大的香港上市内地企业股票构成，部分数据见表 1。指数每半年检讨一次，新上市股票可快速加入指数；而指数使用自由流通量市值进行加权计算，反映了市场实际可投资份额。参考 2012 年 10 月 31 日的收市价，中华 120 指数流通市值为 61,660 亿元人民币，平均市值为 514 亿元。按行业划分，占指数比重最大的是金融类股份(即银行、保险、地产)，占比为 48%；其次为能源，占比 15%。

Table 1. China 120 index data presentation
表 1. 中华 120 指数部分数据

Time	X1	X2	X3	X4	X5		X120	Y
22/04/11	16.05	20.74	55.36	180.10	31.78		170.50	5815.483
22/04/12	15.92	20.17	57.44	193.00	32.46		171.05	5921.478
22/04/13	15.80	20.5	56.57	192.70	31.96		169.08	5919.987
22/04/14	16.04	21.00	58.72	209.53	32.49		172.43	5987.719
22/04/15	16.42	21.3	58.65	208.50	32.38		173.90	5984.630
22/04/18	15.90	20.53	57.57	206.10	32.00		182.71	5944.459
22/04/19	15.81	20.66	57.30	202.95	31.55		164.44	5840.193
22/04/20	15.85	19.35	57.56	202.91	31.73		159.70	5789.341
22/04/21	15.81	19.30	56.03	199.39	31.18		155.33	5695.934
22/04/22	16.06	19.81	56.97	201.40	31.97		149.00	5730.146
22/04/25	14.85	18.83	55.07	192.70	30.71		138.21	5508.757
.....								
23/04/04	12.65	15.15	54.63	259.23	37.03		95.580	5792.286
23/04/06	12.58	15.04	54.02	250.04	36.75		101.10	5784.381
23/04/07	12.62	15.82	54.68	249.00	37.29		103.52	5803.372

其中变量 X_1 、 X_2 、……、 X_{120} 分别为平安银行、万科 A、美的集团、泸州老窖、格力电器、京东方 A、盐湖股份、五粮液、中国旺旺、吉利汽车、宁波银行、洋河股份、顺丰控股、阿里健康、海康威视、立讯精密、荣盛石化、比亚迪、牧原股份、恩捷股份、华润啤酒、中国海外发展、腾讯控股、恒大汽车、中国联通、中国海洋石油、建设银行、中国移动、龙湖集团、联想集团、快手-W、华润置地、比亚迪股份、工商银行、小米集团-W、碧桂园、理想汽车-W、安踏体育、药明生物、申洲国际、中国平安、蒙牛乳业、李宁、舜宇光学科技、京东物流、交通银行、美团-W、招商银行、中国银行、碧桂园服务、京东健康、海底捞、思摩尔国际、京东集团-SW、农夫山泉、小鹏汽车-W、百度集团-SW、网易-S、亿纬锂能、爱尔眼科、东方财富、智飞生物、阳光电源、宁德时代、迈瑞医疗、金龙鱼、浦发银行、民生银行、宝钢股份、中国石化、中信证券、三一重工、招商银行、保利发展、中国联通、上汽集团、恒瑞医药、万华化学、恒力石化、国电南瑞、片仔癀、通威股份、贵州茅台、海螺水泥、海尔智家、山西汾酒、伊利股份、长江电力、三峡能源、招商证券、隆基绿能、中信建投、中国神华、工业富联、兴业银行、国泰君安、农业银行、中国平安、中国人保、交通银行、工商银行、中国太保、中国人寿、长城汽车、邮储银行、中国建筑、中国电信、中国中车、京沪高铁、光大银行、中国石油、中国中免、紫金矿业、中远海控、中国银行、中金公司、中信银行、药明康德、海天味业、韦尔股份。

3.2. 数据集划分

本文 120 个变量，使用 2/3 作为训练集，1/3 作为测试集，见表 2。

Table 2. Data set segmentation

表 2. 数据集划分

数据集	训练集	测试集
$(X, Y)_{243 \times 121}$	$(X, Y)_{162 \times 121}$	$(X, Y)_{81 \times 121}$

介绍几种常用的判断追踪能力的方法。追踪偏差(简称为 TD_i)是指追踪组合的日收盘价与中华 120 指数的日收盘价之间的偏差，一般都用以下公式计算：

$$TD_i = y_i - \hat{y}_i \quad (11)$$

本文判断追踪能力的方法具体描述如下：残差平方和、平均残差平方和、残差标准差

$$\text{残差平方和: } S_E^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

$$\text{平均残差平方和: } RMS = \frac{S_E^2}{n-p} \quad (13)$$

$$\text{残差标准差: } SD = \sqrt{\frac{(\text{error} - \overline{\text{error}})^2}{n-1}} \quad (14)$$

其中， $\hat{y}_i$ 是指追踪组合的日收盘价， y_i 中华 120 指数的日收盘价。

4. 指数追踪

4.1. 逐步回归

通过 R 中函数 `step()`，选择 AIC 信息量为准则，默认向后法，从所有变量开始，逐步通过选择最小的 AIC 信息量达到增删变量的目的。首先，我们通过逐步回归，得到结果显示剔除 45 个变量(X_7 , X_8 ,

$X_9, X_{11}, X_{14}, X_{19}, X_{20}, X_{25}, X_{36}, X_{38}, X_{39}, X_{46}, X_{47}, X_{48}, X_{50}, X_{51}, X_{54}, X_{55}, X_{60}, X_{62}, X_{63}, X_{64}, X_{67}, X_{70}, X_{71}, X_{74}, X_{78}, X_{81}, X_{82}, X_{88}, X_{90}, X_{93}, X_{95}, X_{98}, X_{99}, X_{103}, X_{104}, X_{106}$), 剩余 75 个变量, 分别为平安银行、万科 A、美的集团、泸州老窖、格力电器、京东方 A、宁波银行、顺丰控股、北方华创、晶澳科技、立讯精密、荣盛石化、比亚迪、华润啤酒、中国海外发展、腾讯控股、中国联通、建设银行、中国移动、龙湖集团、快手-W、华润置地、比亚迪股份、工商银行、小米集团-W、碧桂园、理想汽车-W、中通快递-SW、中国平安、蒙牛乳业、李宁、舜宇光学科技、京东物流、美团-W、百济神州、思摩尔国际、京东集团-SW、百度集团-SW、百胜中国、阿里巴巴-SW、网易-S、爱尔眼科、阳光电源、宁德时代、金龙鱼、浦发银行、中国石化、中信证券、招商银行、保利发展、上汽集团、万华化学、恒力石化、通威股份、贵州茅台、海螺水泥、海尔智家、山西汾酒、长江电力、隆基绿能、中信建投、工业富联、陕西煤业、农业银行、交通银行、工商银行、中国太保、邮储银行、中国电信、中国中车、京沪高铁、光大银行、中国石油、中国中免、紫金矿业、中远海控、中国银行、中金公司、中信银行、药明康德、海天味业、韦尔股份。

然后对数据中华 120 指数进行参数估计。对于 X_2, X_{10}, X_{30} 等几个变量, 虽然变量不显著, 但删除后 AIC 和残差反而增加, 因此我们对这几个变量进行保留。在该模型下对应的拟合残差图、预测残差图、中华 120 指数预测图、中华 120 指数跟踪图, 见图 1。

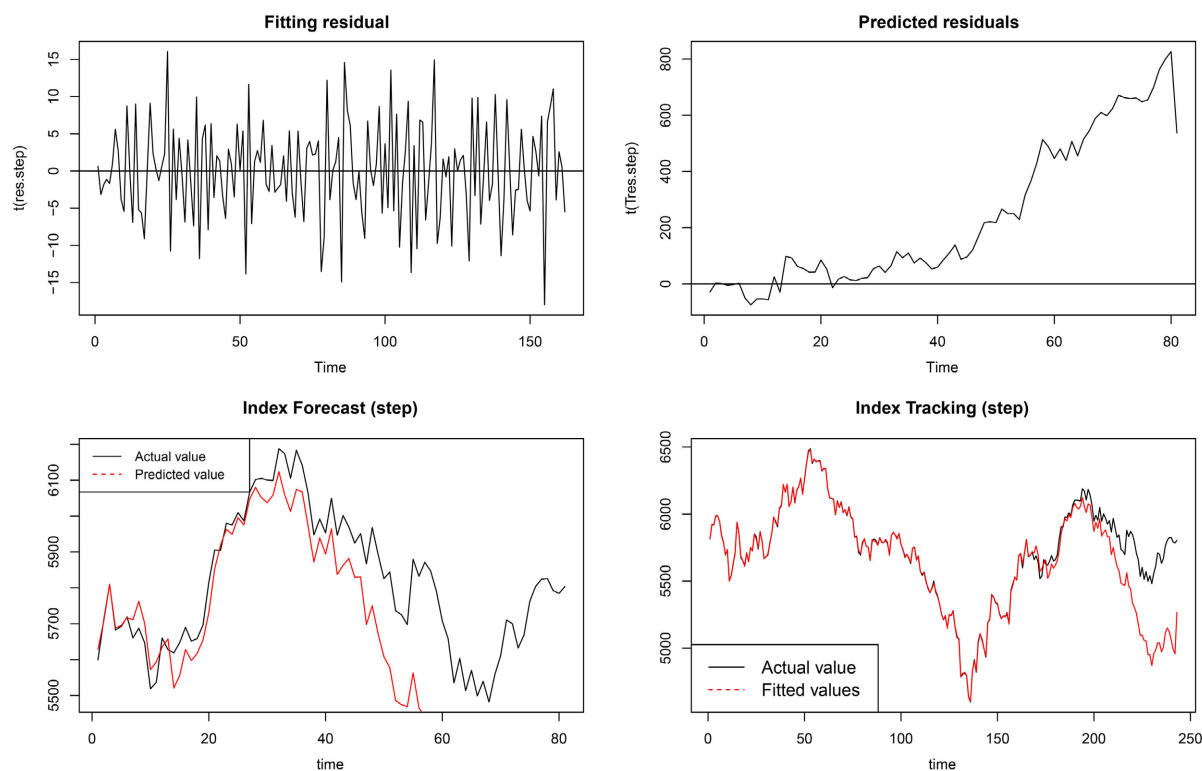


Figure 1. Fitted residuals, predicted residuals, index predictions, index tracking charts

图 1. 拟合残差、预测残差、指数预测、指数跟踪图

4.2. Lasso

4.2.1. 绝对约束估计

对于绝对约束估计, 我们考虑两个问题: 一是否存在最优子集? 二是如何选出最优子集? 当然, 这

样的子集包含的变量越少越好。

一般认为，权重股由于盘大，往往能起到四两拨千斤的作用，但正是因为盘大，走势呆滞，往往对指数起伏贡献甚微。所以合理的选择股票非常重要，我们可以用变量选择来完成这份工作。传统的方法要丢弃较多变量比较困难，而 Lasso 具有的稀疏性恰好在此时可以派上用场。

下面先用 LARS 的值选择模型，见图 2。在 C_p 准则下，选择最小的 C_p 值对应的变量集，结果显示，最小值 $C_p = 100.1821$ ，对应的变量集包含 89 个变量。即通过变量选择，从原始 120 个变量选择了 89 个变量进行指数追踪。

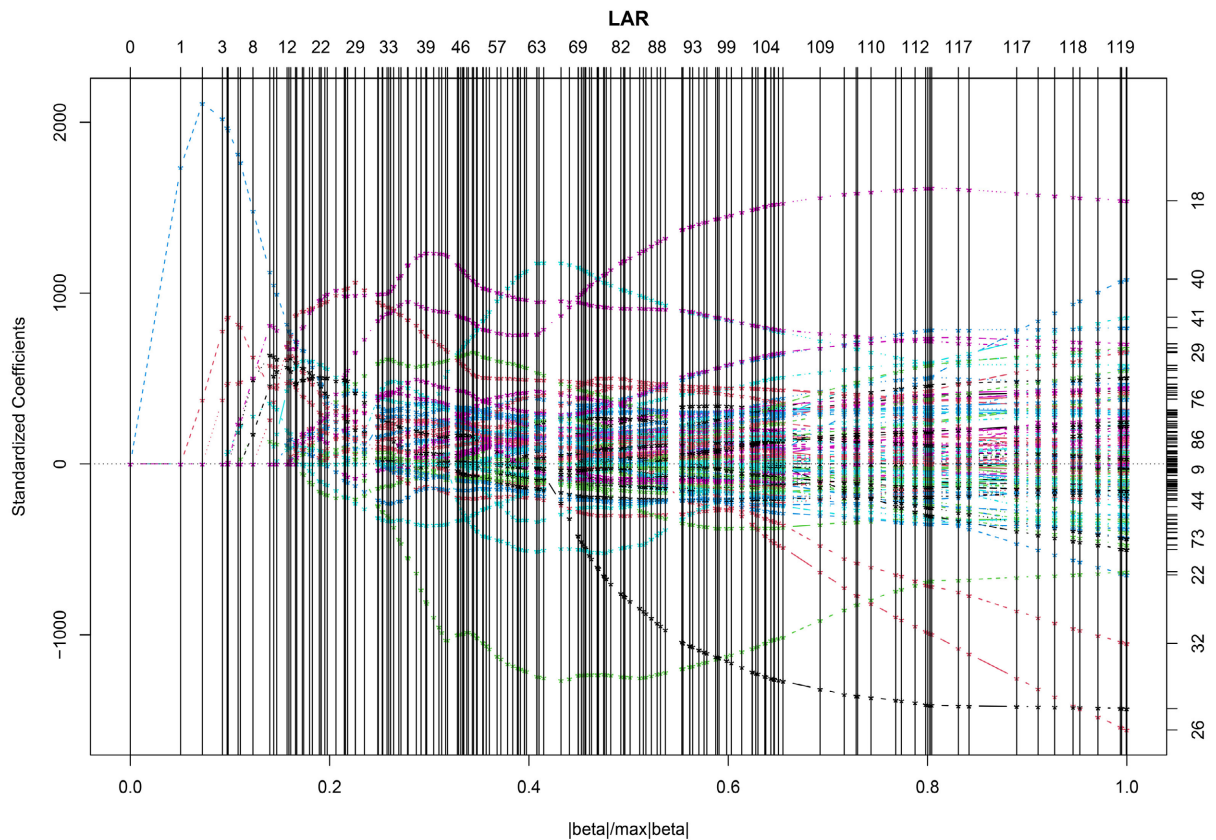


Figure 2. Diagram of the coefficients for each group of variables

图 2. 各组变量系数示意图

在该模型下对应的拟合残差图、预测残差图、中华 120 指数预测图、中华 120 指数跟踪图见图 3。

4.2.2. 弹性约束估计

虽然绝对约束估计在很多情况下都得到很大的认可，但有效性在某些条件下也会受到限制，主要在如下三个方面：

- 1) 在 $p > n$ 的情况下，绝对约束估计最多只能选择出 n 个变量。
- 2) 在一组相关性较高的变量中，绝对约束估计只能在这些变量中选择其中的一个，而不考虑其他具有较高相关性的变量，选择也是随意的；
- 3) 就低维情形， $p < n$ 的情况下，如果预测值之间有较高的相关性，那么岭回归估计比绝对约束估计表现要好。

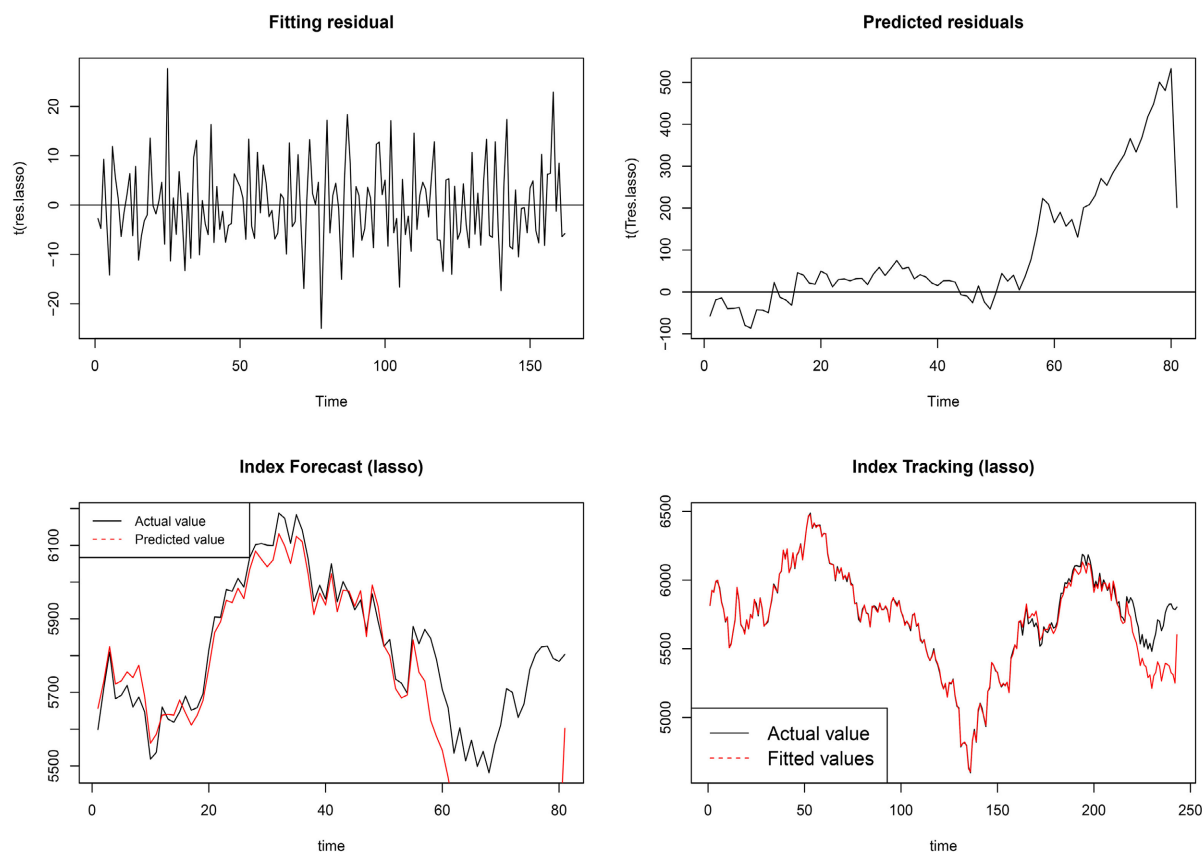


Figure 3. Fitted residuals, predicted residuals, index predictions, index tracking charts

图 3. 拟合残差、预测残差、指数预测、指数跟踪图

2005 年 Zou, H 和 Hastie, T 提出合并考虑岭回归和 LASSO 的约束方式, 提出了弹性约束估计, 称之为 Elasticnet, 定义如下:

$$\tilde{\beta}_j = \arg \min_{\beta} \left(\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right) \quad (15)$$

等价于找到使得下式

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2, \text{ 满足 } (1-\lambda) \sum_{j=1}^p |\beta_j| + \lambda \sum_{j=1}^p \beta_j^2 \leq t \quad (16)$$

达到最小的 $\beta_j, j=1, 2, \dots, p$

其实, 当 $\lambda=1$ 时弹性约束估计就是岭回归, 当 $\lambda=0$ 弹性约束估计就是绝对约束估计, 因此, 弹性约束估计同时具有绝对约束估计和岭估计的特点。R 软件中 glmnet 程序包可用于计算弹性约束估计。

在这里, 我们先用弹性约束估计建立中华 120 指数与成分股的回归方程。比较方便的是函数 cv.glmnet 可以自动进行 CV 交叉验证, 从而确定出最佳的值, 对于本文来说, $\lambda_{\min} = 0.2503$, λ 值的选择从图 4(上)中一目了然, 同时也给出回归系数图, 如图 4(下)。

上图中的横轴是对数 λ 值, 纵轴是均方误差。按此参数值, 保留变量包含 72 个变量。在该模型下对应的拟合残差图、预测残差图、中华 120 指数预测图、中华 120 指数跟踪图如图 5。

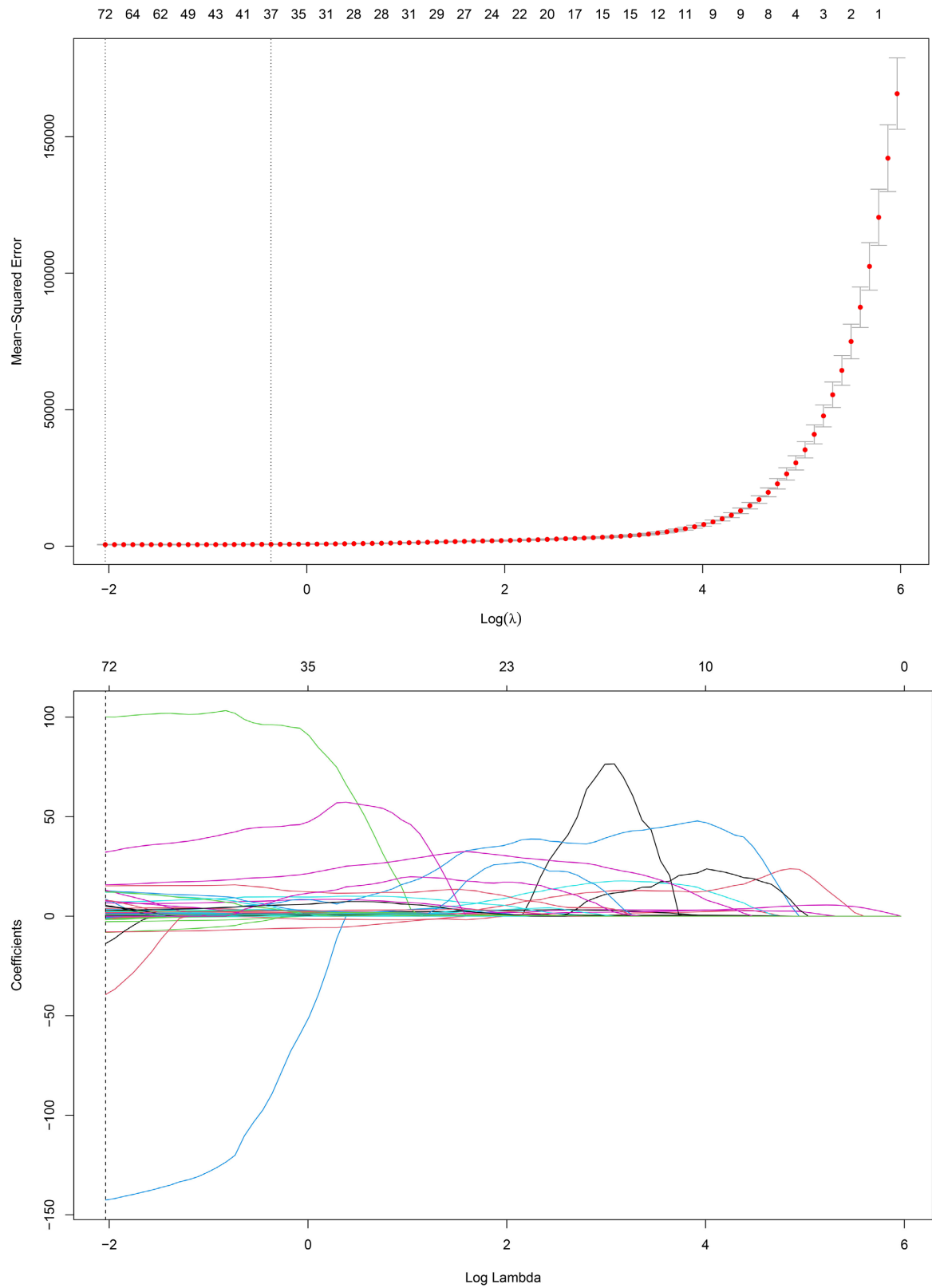


Figure 4. λ choice plot (Top), regression coefficient plot (Bottom)

图 4. λ 选择图(上)、回归系数图(下)

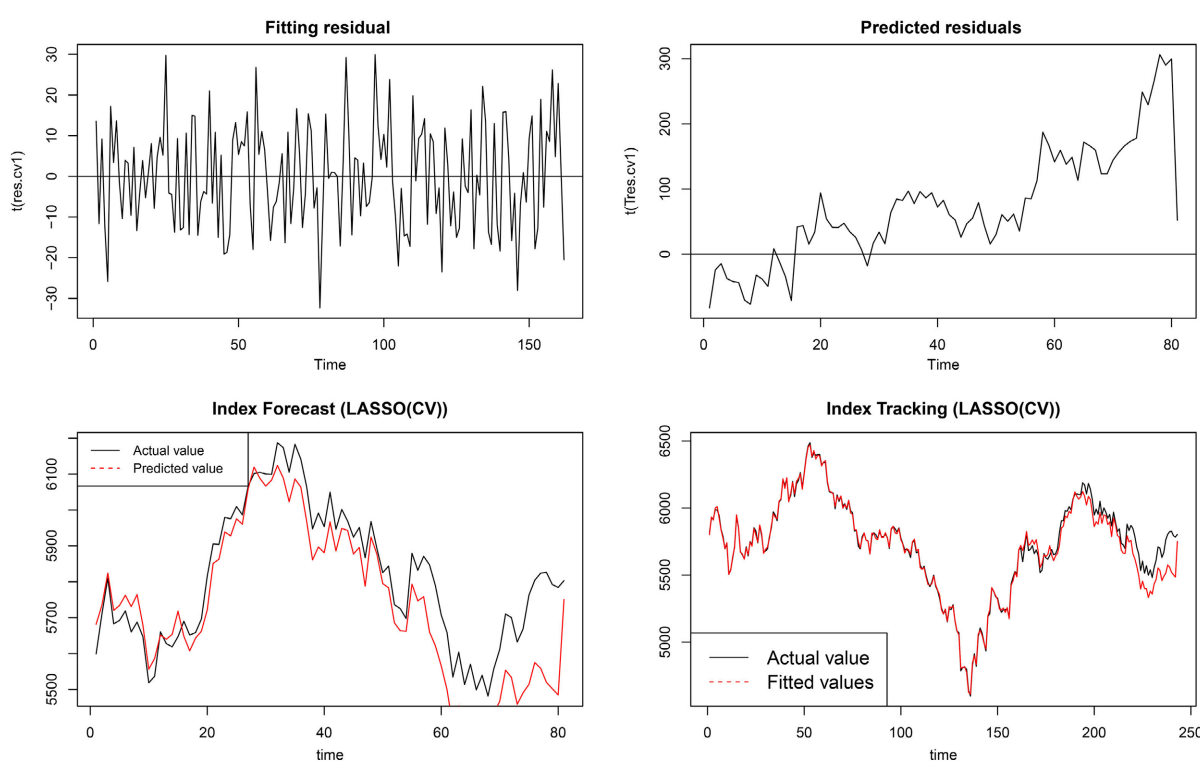


Figure 5. Fitted residuals, predicted residuals, index predictions, index tracking charts
图 5. 拟合残差、预测残差、指数预测、指数跟踪图

4.3. 岭回归

岭回归先于 Lasso 方法出现，但岭回归建立的是 L2 正则化模型，Lasso 建立的是 L1 正则化模型。两者区别在于，岭回归只会使模型参数趋近于 0 而不是等于 0，即模型训练完成后每一个参数都会有一个非零权值；而 Lasso 回归会使大部分模型参数等于 0 以实现稀疏化。与 Lasso 相比，岭回归得到的模型一直都是 120 个变量，因此岭回归没有变量筛选的功能。因此，运用 Lasso 进行变量选择，然后再通过最小二乘等回归方法对筛选出来的变量进行回归，从而对指数进行跟踪。

本文利用 R 语言中的 MASS 包、glmnet 包实现岭回归。通过 `cvfit2$lambda.min`、`cvfit2$lambda.1se` 得到均方误差最小时的 λ 值为 38.93432、距离均方误差最小时一个标准误的 λ 值为 63.08443，同时进行交叉验证并绘图，如图 6(上)。

图 6(下)中横坐标为随着 λ 的对数，纵坐标为变量系数，从图中我们可以看到随着 λ 增加变量系数不断减少，部分变量系数变为 0 (等于没有这个变量了)。岭回归将系数缩小为零，但不会将其中的任何一个都精确设置为零。在该模型下对应的拟合残差图、预测残差图、中华 120 指数预测图、中华 120 指数跟踪图如图 7。

4.4. 效果对比

从上面基于不同方法的指数追踪效果来看，似乎无法辨别某种方法下的中华 120 指数的追踪效果最佳，某种方法下的中华 120 指数的追踪效果最差，这个时候就需要用数据、用评价标准来判别基于不同方法下指数追踪效果的优劣了。在表 3 和表 4 中，我们展示了基于不同准则下(C_p 准则，CV 准则)的方法的追踪效果，以及在同一准则下不同方法的追踪效果。

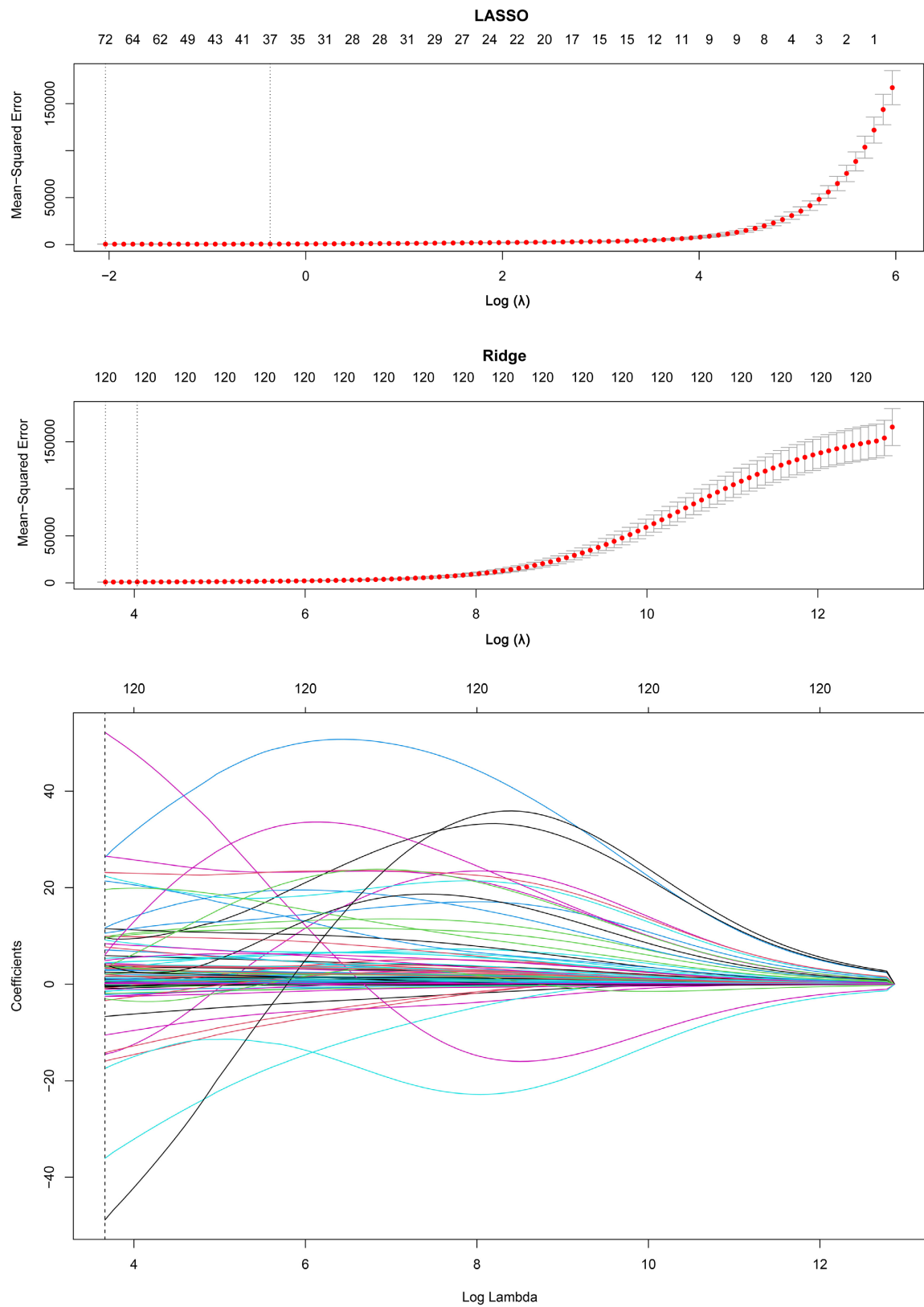


Figure 6. Cross-validation plot (Top), ridge regression coefficient plot (Bottom)

图 6. 交叉验证图(上)、岭回归系数图(下)

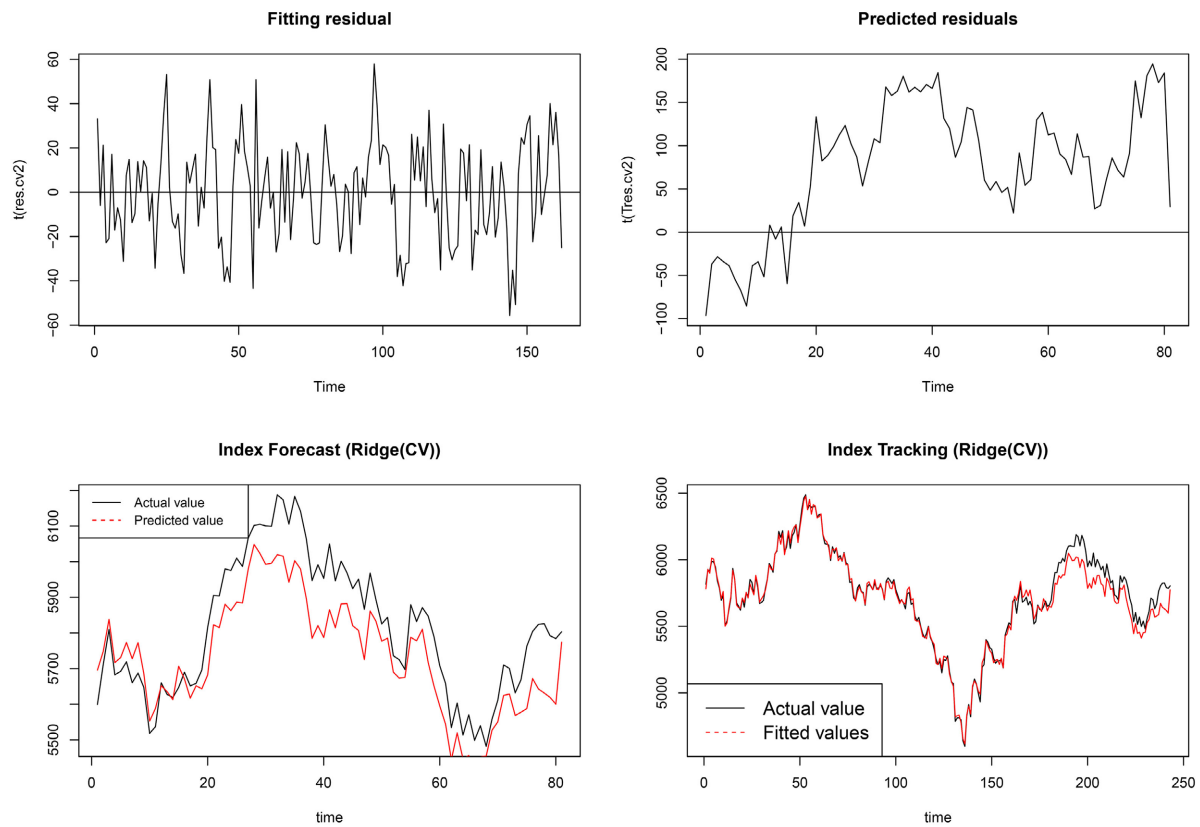


Figure 7. Fitted residuals, predicted residuals, index predictions, index tracking charts
图 7. 拟合残差、预测残差、指数预测、指数跟踪图

Table 3. Comparison of the effectiveness of different methods of index tracking under the C_p guidelines

表 3. C_p 准则下不同方法的指数追踪效果对比

方法(C_p 准则)	训练集			测试集		
	SSE	RMS	SD	SSE	RMS	SD
逐步回归(75 个变量)	5386.56	71.82	6.3220	1,689,352	22524.70	72.9586
Lasso (89 个变量)	12506.73	140.53	6.0550	1,771,612	19905.75	26.0402
岭回归	146418.91	2033.60	22.7600	2,376,360	20803.24	27.5321
Lasso + olse	10888.67	145.18	6.5800	2,715,301	36204.01	25.7395
Lasso + 岭回归	16734.67	223.28	6.7650	1,963,476	26179.68	22.5828
Lasso + LIU (olse)	15244.06	203.22	7.7840	1,692,975	22572.92	19.9017

Table 4. Comparison of the effectiveness of different methods of index tracking under the CV guidelines

表 4. CV 准则下不同方法的指数追踪效果对比

方法(CV 准则)	训练集			测试集		
	SSE	RMS	SD	SSE	RMS	SD
逐步回归(75 个变量)	5386.56	71.82	6.3221	1,689,352	22524.70	72.9586

Continued

Lasso (72 个变量)	11828.93	164.29	8.5716	1,086,600	15091.68	21.5125
岭回归	146418.91	2033.60	22.7568	2,376,360	19803.24	24.5321
Lasso + olse	9447.08	78.73	6.4847	1,433,214	19905.75	26.0402
Lasso + 岭回归	100158.56	1391.09	6.5728	1,368,249	19003.47	25.6692
Lasso + LIU (olse)	8125.61	112.86	7.7642	868523.8	12062.85	20.4668

在 C_p 准则下, Lasso 保留了 89 个变量(成分股), 逐步回归保留 75 个变量。且在测试集上 Lasso 的残差平方和、平均残差平方和(RMS)和残差标准差(SD)三种指标都优于逐步回归和岭回归; 在 Lasso 变量选择方法下, 进一步运用 LIU 估计进行回归, 得到较好的外预测效果。

相比于 C_p 准则, CV 准则下的 Lasso 只保留了 72 个变量, 逐步回归保留 75 个变量。外预测效果明显都次于 C_p 准则; 在 CV 准则下, Lasso 测试集上的残差平方和、平均残差平方和(RMS)和残差标准差(SD)三种指标都优于逐步回归和岭回归; 两步估计中, LIU 估计外预测效果也是较好的。

5. 结论

随着国内外证券市场的成熟化, 追踪标的国际化, 投资者的指数化投资理念更加成熟, 越来越多的指数型基金管理公司开始通过构建指数追踪组合对市场指数进行追踪。因此, 本文研究的指数追踪方法对国内外资本市场都具有十分重要的理论意义和实际意义, 也为投资者提供更多的方法。

本文的主要内容是通过准则、CV 准则选股法来选出构建追踪组合的样本股, 并且分别构建了指数追踪的逐步回归模型、Lasso、岭回归以及两步回归模型。

本文得到的主要结论如下:

- 1) 在 C_p 准则下, Lasso 保留了 89 个变量(成分股), 逐步回归保留 75 个变量。且在测试集上的残差平方和、平均残差平方和(RMS)和残差标准差(SD)三种指标 Lasso 都优于逐步回归和岭回归; 在 Lasso 变量选择方法下, 进一步运用 LIU 估计进行回归, 得到较好的外预测效果。
- 2) 在 CV 准则下, Lasso 只保留了 72 个变量, 逐步回归保留 75 个变量。外预测效果明显都次于准则; 在 CV 准则下, Lasso 测试集上的残差平方和、平均残差平方和(RMS)和残差标准差(SD)三种指标都优于逐步回归和岭回归; 两步估计中, LIU 估计外预测效果也是较好的。

参考文献

- [1] 杨国梁, 赵社涛, 徐成贤. 基于支持向量机的金融市场指数追踪技术研究[J]. 国际金融研究, 2009, 270(10): 68-72.
- [2] 李俭富, 马永开, 曾勇. 指数跟踪研究评述[J]. 当代经济管理, 2005(6): 19-24.
- [3] 李俭富, 马永开, 曾勇. 指数跟踪组合复制方法的实证研究[J]. 管理学报, 2006(3): 354-359+378.
- [4] Canakgoz, N.A. and Beasley, J.E. (2009) Mixed-Integer Programming Approaches for Index Tracking and Enhanced Indexation. *European Journal of Operational Research*, **196**, 384-399. <https://doi.org/10.1016/j.ejor.2008.03.015>
- [5] 王玉玲, 马军海, 王晶. 主成分分析法在股票研究中的应用[J]. 现代管理科学, 2009, 8(11): 63-65.
- [6] 陈春锋, 陈伟忠. 指数优化复制的方法、模型与实证[J]. 数量经济技术经济研究, 2004(12): 106-115.
- [7] 陈伟忠, 李健飞, 陈春锋. 指数优化复制中的流动性改进——基于上证 180 指数的实证研究[J]. 系统工程, 2005(2): 55-61.
- [8] 顾明庆, 程相怡. 基于 LASSO 算法的股指追踪研究[J]. 投资与合作, 2023, 387(2): 42-44.

- [9] 王元, 文兰, 陈木法. 数学大辞典[M]. 北京: 科学出版社, 2010.
- [10] 张明芝, 李红美, 吕大兵, 主编. 汤在祥, 裴育芳, 副主编. 实用医学统计学与 SAS 应用[M]. 苏州: 苏州大学出版社, 2015.