Operations Research and Fuzziology 运筹与模糊学, 2023, 13(6), 7741-7756
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136757

追踪市场指数的投资组合策略

——以上证50指数为例

叶茂越

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2023年9月22日；录用日期：2023年12月22日；发布日期：2023年12月29日

摘　要

指数追踪是一种用少量的成分股来追踪某一市场指数走势的方法，它是消极投资组合管理策略中的一种。本文通过逐步回归、岭回归、两步回归和分位数回归等构建指数追踪模型，并通过C_p准则、CV准则得到了两个样本股空间，在这两个样本股空间上对模型进行实证分析。本文选取了2021年8月1日到2022年7月1日的上证50指数的日线收盘价数据，划分2/3训练集和1/3测试集，最后通过残差平方和、平均残差平方和、残差标准差等评价指标进行对比分析。本文得到的主要结论如下：1) 在Cp准则下，LASSO保留了49个变量(成分股)，且在测试集上的残差平方和、平均残差平方和和残差标准差三种指标都优于逐步回归和岭回归；在LASSO变量选择方法下，进一步运用刘估计进行回归，得到较好的外预测效果。2) 在CV准则下，LASSO只保留了36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、平均残差平方和和残差标准差三种指标都优于逐步回归和岭回归；两步估计中，岭估计外预测效果也是较好的。3) 在CV准则下，LASSO测试集上的残差平方和、平均残差平方和和残差标准差三种指标都优于逐步回归和岭回归；两步估计中，0.5分位数回归外预测效果也是最好的，0.05分位数回归的效果最差。4) 在不同的选股法下数值的改变对模型的影响不同。

关键词

指数追踪，两步回归，分位数回归，上证50指数

A Portfolio Strategy for Tracks Market Indices

—Taking the SSE 50 Index as an Example

Maoyue Ye

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 22nd, 2023; accepted: Dec. 22nd, 2023; published: Dec. 29th, 2023

文章引用: 叶茂越. 追踪市场指数的投资组合策略[J]. 运筹与模糊学, 2023, 13(6): 7741-7756.
DOI: 10.12677/orf.2023.136757

Abstract

Index tracking is a method of tracking the movement of a market index with a small number of constituent stocks. It is one of the passive portfolio management strategies. In this paper, the index tracking model is constructed by stepwise regression, ridge regression, two-step regression and quantile regression, and two sample stock Spaces are obtained by C_p criteria and CV criteria, and empirical analysis is conducted on these two sample stock Spaces. This paper selects the daily closing price data of the Shanghai Stock Exchange 50 Index from August 1, 2021 to July 1, 2022, divides the 2/3 training set and 1/3 test set, finally, the evaluation indexes such as sum of squares of residuals, average sum of squares of residuals and standard deviation of residuals are compared and analyzed. The main conclusions of this paper are as follows: 1) under the C_p criteria, 49 variables (component stocks) are retained in LASSO, and the three indexes of residual sum of squares, mean residual sum of squares and residual standard deviation on the test set are better than stepwise regression and ridge regression; under the LASSO variable selection method, Liu estimation is further used for regression, and better external prediction results are obtained. 2) Under the CV criterion, LASSO retains only 36 variables, and the external prediction effect is obviously inferior to the C_p criterion; under the CV criterion, the sum of squares of residuals, sum of squares of mean residuals and standard deviation of residuals on LASSO test set are better than stepwise regression and ridge regression. In the two-step estimation, the prediction effect outside the ridge estimation is also better. 3) Under CV criteria, the three indexes of the sum of squares of residuals, the sum of squares of mean residuals and the standard deviation of residuals on LASSO test set are better than stepwise regression and ridge regression; in the two-step estimation, the effect of 0.5 quantile regression is the best, and the effect of 0.05 quantile regression is the worst. 4) The change of values under different stock selection methods has different impacts on the model.

Keywords

Index Tracking, Two-Step Regression, Quantile Regression, SSE 50 Index

1. 引言

随着股票价格指数的发展与演变，指数型衍生品创新日益成为了当今全球金融市场上的一大亮点。在金融危机中表现突出，被大多数投资者和机构所采用，如何用部分股票来追踪目标指数就成为是否采用指数法进行投资的关键所在。股票指数追踪研究的内容是通过权重的优化再配置来寻找部分股票构成的最优的追踪组合，所谓的最优就是使得该组合相对标的指数的追踪误差最小。其目的在于复制与该指数同样收益水平的一个投资组合，实现组合收益与指数涨跌基本一致. 此研究具有高精度、低交易成本、且能保证追踪组合的高流动性，具有重要的意义。

目前已有不少学者对此进行了研究，Roll (1992) [1]以经典的均值—方差模型为切入点采用追踪误差最小化为目标函数。Alexander 和 Baptista (2010) [2]提出了追踪边缘的曲率。Gotoh 和 Takeda (2011) [3]就范数约束和惩罚对投资组合选择问题进行研究，研究发现范数约束可被看作与收益向量有关的稳健约束。梁斌等(2011) [4]提出非负 LASSO 模型用以追踪指数，实证结果表明非负 LASSO 模型表现较好。刘睿智和周勇(2015) [5]将自适应 LASSO 变量选择方法运用于指数追踪模型中并得到较好追踪效果。苏治

等(2016) [6]基于规则化方法构建稀疏指数追踪模型具有较好的外推预测效果。胡梦婷(2017) [7]提出了指数追踪的分位数回归模型，并得到结论在不同的选股法下K值的改变对模型的影响不同。马景义(2017) [8]构建了可以调节追踪误精和超额收益的增强型指数追踪模型，并给出了广义最小角度回归算法。用以计算调节参数作用下模型解的折中路径。彭胜银(2019) [9]指出非负分位数估计作为两步估计方法，不仅保证了追踪组合内的股票权重的非负性；同时，还可以自动剔除一些相关性较高的成分股，降低追踪误差，提高追踪组合的样本内外预测能力，减少交易成本，说明该方法在股指追踪上的有效性和可行性。

本文在解决股票指数追踪问题时，先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。本文将以上证50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. 数据说明

数据选取：2021年8月1日到2022年7月1日的上证50指数的日线收盘价数据，共207组样本，我们将数据集进行划分：2/3训练集，则训练集数据有138个样本，1/3测试集，则测试集数据有69个样本。表1为上证50指数上证(000016.SH)的50只成分股。

Table 1. The 50 constituent stocks of the SSE 50 index (000016.SH)
表1. 上证50指数上证(000016.SH)的50只成分股

变量	名称	代码	变量	名称	代码	变量	名称	代码
X1	包钢股份	600010	X18	用友网络	600588	X35	中国太保	601601
X2	中国石化	600028	X19	海尔智家	600690	X36	中国人寿	601628
X3	中信证券	600030	X20	闻泰科技	600745	X37	长城汽车	601633
X4	三一重工	600031	X21	山西汾酒	600809	X38	中国建筑	601668
X5	招商银行	600036	X22	海通证券	600837	X39	华泰证券	601688
X6	保利发展	600048	X23	伊利股份	600887	X40	中国电信	601728
X7	上汽集团	600104	X24	航发动力	600893	X41	中国石油	601857
X8	北方稀土	600111	X25	长江电力	600900	X42	中国中免	601888
X9	复星医药	600196	X26	三峡能源	600905	X43	紫金矿业	601899
X10	恒瑞医药	600276	X27	隆基绿能	601012	X44	中远海控	601919
X11	万华化学	600309	X28	中信建投	601066	X45	中金公司	601995
X12	恒力石化	600346	X29	中国神华	601088	X46	药明康德	603259
X13	片仔癀	600436	X30	兴业银行	601166	X47	海天味业	603288
X14	通威股份	600438	X31	国泰君安	601211	X48	韦尔股份	603501
X15	贵州茅台	600519	X32	农业银行	601288	X49	华友钴业	603799
X16	恒生电子	600570	X33	中国平安	601318	X50	兆易创新	603986
X17	海螺水泥	600585	X34	工商银行	601398			

首先我们先介绍一下追踪偏差(Tracking deviation，简写为TD_i)的概念，是指追踪组合的日收盘价与上证50指数的日收盘价之间的偏差，一般都用以下公式计算：

$$TD_i = y_i - \hat{y}_i \tag{1}$$

本文判断追踪能力的方法具体描述如下：

残差平方和(Residual Sum of Squares)： $S_E^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$

平均残差平方和(Mean Residual Sum of Squares)： $\mathrm{RMS}=\frac{S_E^2}{n-p}$

残差标准差(Residual Standard Deviation)： $\mathrm{SD}=\sqrt{\frac{\left(\mathrm{error}-\overline{\mathrm{error}}\right)^2}{n-1}}$

其中， $\hat{y}_i$ 是指追踪组合的日收盘价， y_i 上证 50 指数的日收盘价。

3. 逐步回归指数追踪模型

逐步回归是通过剔除变量中不太重要又和其他变量高度相关的变量，降低多重共线性程度。将变量逐个引入模型，每引入一个解释变量后都要进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除，以确保每次引入新的变量之前回归方程中只包含显著性变量。这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止，以保证最后所得到的解释变量集是最优的。

逐步回归法的好处是将统计上不显著的解释变量剔除，最后保留在模型中的解释变量之间多重共线性不明显，而且对被解释变量有较好的解释贡献。但逐步回归法可能因为删除了重要的相关变量而导致设定偏误。

逐步回归的方法有两种，一种是向前法，另一种是向后法，本文通过采用函数 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。首先通过逐步回归剔除 7 个变量 $\left(X_8,X_{29},X_{31},X_{36},X_{39},X_{45},X_{49}\right)$，余下 43 个变量，对于第 $\left(X_{12},X_{42}\right)$ 个变量，虽然不显著，但删除后 AIC 和平均残差平方和、残差标准差等有所增加，因此不删除。所选择的变量给出的参数估计，得到回归模型如下：

$$\begin{aligned}\hat{Y}_1=&-90.79+7.07X_1+13.90X_2+3.20X_3+2.73X_4+5.55X_5+3.84X_6+1.61X_7+0.34X_9+2.31X_{10}\\&+0.61X_{11}+0.35X_{12}+0.08X_{13}+1.09X_{14}+0.27X_{15}+0.68X_{16}+0.85X_{17}+1.03X_{18}+1.30X_{19}\\&+0.27X_{20}+0.15X_{21}+6.36X_{22}+2.14X_{23}+0.55X_{24}+3.82X_{25}-3.14X_{26}+1.98X_{27}+1.60X_{28}\\&+5.18X_{30}+39.00X_{32}+3.66X_{33}+36.23X_{34}+3.96X_{35}-0.23X_{37}+6.97X_{38}+3.24X_{40}+3.31X_{41}\\&+0.50X_{42}+10.08X_{43}+0.80X_{44}+0.68X_{46}+0.34X_{47}+0.29X_{48}+0.30X_{50}\end{aligned}\tag{2}$$

在 $\hat{Y}_1$ 模型下对应的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图，如下图 1 所示。

从图 1 中的拟合残差图可知，该拟合残差在 0 上下波动，无明显趋势效应；从预测残差图显示，不是该残差白噪声序列；上证 50 指数预测图可以看出随着时间的增加，预测误差越大，预测值显著低于实际值，可以看出预测效果不好；上证 50 指数跟踪图可以看出跟踪效果不佳。综上所述，此时在 $\hat{Y}_1$ 模型下的指数追踪效果并不好。

4. LASSO 回归指数追踪模型

考虑线性模型 $y=\beta_0 I+X\beta+\varepsilon$，误差向量 ε 满足 $E\left(\varepsilon\right)=0,Var\left(\varepsilon\right)=\sigma^2$，样本数据为 $\left(x_i,y_i\right),i=1,2,\cdots,n$，其中， $x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)$ 是自变量， y_i 是因变量，假设这些观测值是相互独立的，或者在给定 x_{ij} 的情况下， y_i 是独立的，所有 x_{ij} 是标准化了。

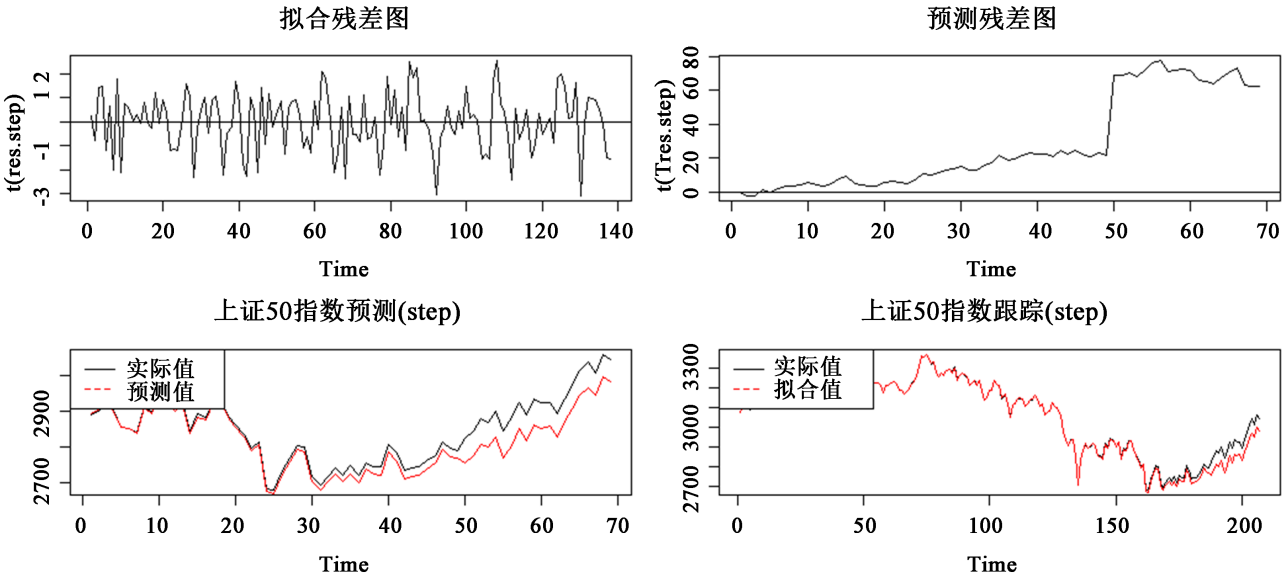

Figure 1. $\hat{Y}_1$ model of fitting residual graph, forecast residual graph, SSE 50 index forecast graph, SSE 50 index tracking graph

图 1. $\hat{Y}_1$ 模型的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图

设 $\hat{\beta}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)$，由 $\left(y-X\beta\right)'\left(y-X\beta\right)+k\beta'\beta$，岭估计其实就是使下式达到最小的参数估计

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+k\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

因此，岭估计可以看成是具有二次约束的回归参数估计，注意，当数据是标准化的时候有 $\bar{y}=0$，又 β_0 的解为 $\hat{\beta}_0=\bar{y}=0$，所以式 $\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+k\sum_{j=1}^{p}\beta_j^2$ 中不含常数项，等同于使下面的式子最小化

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2,\text{满足}\sum_{j=1}^{p}\beta_j^2\le t \tag{4}$$

绝对约束估计(LASSO: the least absolute shrinkage and selection operator)就是要找到使下式

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2,\text{满足}\sum_{j=1}^{p}\beta_j^2\le t \tag{5}$$

达到最小的 $\beta_j, j=1,2,\cdots,p$，$t\ge 0$ 是一个调整参数，它控制着对估计压缩的程度，设 $\hat{\beta}_j$ 是普通最小二乘估计，$t_0=\sum\left|\hat{\beta}_j\right|$，当 $t\le t_0$，将会引发趋向 0 的压缩，因而绝对约束估计也是压缩估计。

总所周知，传统的方法要丢弃如此众多的变量非常困难。而绝对约束估计(LASSO)具有的稀疏性。下面先用 LARS 的 C_p 值选择模型，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 $C_p=49.71793$，对应的变量集包含 49 个变量，即通过变量选择，从原始 50 个变量选择了 49 个变量进行指数追踪。运用 plot.lars 可以直观的显示各组变量系数，得到回归模型如下：

$$\begin{aligned}\hat{Y}_2=&8.00X_1+11.95X_2+2.94X_3+2.88X_4+5.57X_5+3.79X_6+1.57X_7-0.15X_8+0.33X_9+2.45X_{10}\\&+0.63X_{11}+0.26X_{12}+0.07X_{13}+1.02X_{14}+0.27X_{15}+0.76X_{16}+0.76X_{17}+1.02X_{18}+1.36X_{19}+0.25X_{20}\end{aligned}$$

$$+0.16X_{21}+5.12X_{22}+2.33X_{23}+0.56X_{24}+3.76X_{25}-3.10X_{26}+1.98X_{27}+1.57X_{28}+0.34X_{29}+5.45X_{30}$$
$$+0X_{31}+40.52X_{32}+3.64X33+32.17X_{34}+4.00X_{35}+0.28X_{36}-0.11X_{37}+7.05X_{38}+0.33X_{39}+2.16X_{40} \quad (6)$$
$$+2.76X_{41}+0.50X_{42}+9.82X_{43}+0.83X_{44}+0.06X_{45}+0.66X_{46}+0.31X_{47}+0.28X_{48}+0.05X_{49}+0.32X_{50}$$

从中 $\hat{Y}_2$ 模型可以看出，剔除变量 X_{31}，X_{31} 变量系数被压缩为 0，从而变量选择保留 49 个变量。在 $\hat{Y}_2$ 模型下对应的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图如下图 2 所示。

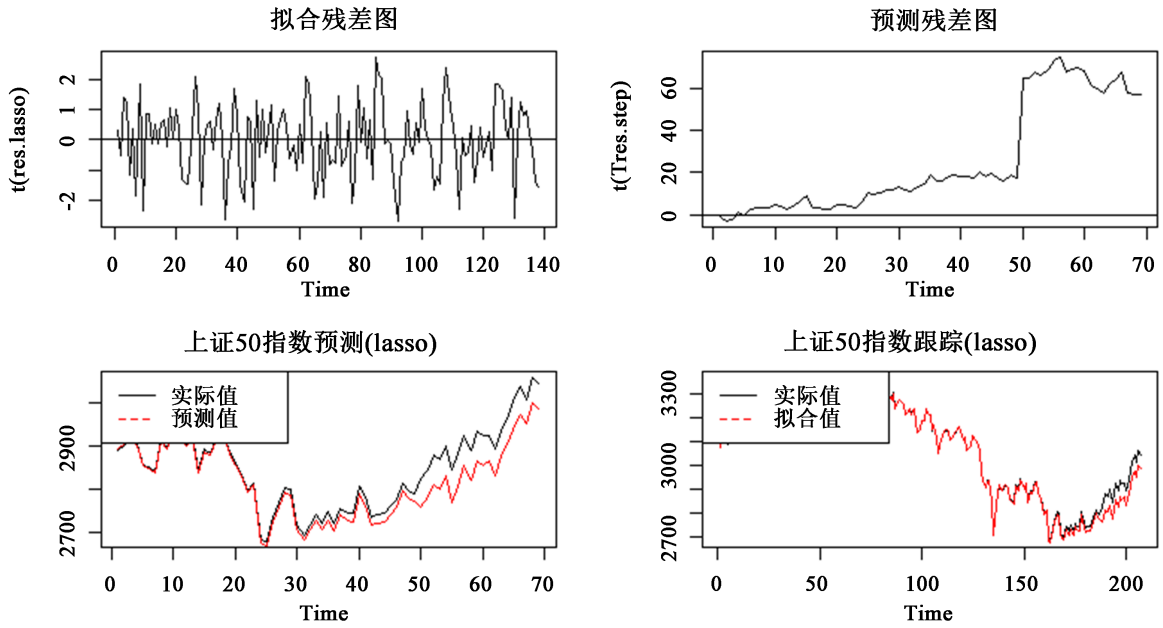

Figure 2. $\hat{Y}_2$ model of fitting residual graph, forecast residual graph, SSE 50 index forecast graph, SSE 50 index tracking graph

图 2. $\hat{Y}_2$ 模型的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图

从图 2 的拟合残差图可知，该拟合残差在 0 上下波动，无明显趋势效应；从预测残差图显示，不是该残差白噪声序列；上证 50 指数预测图可以看出随着时间的增加，预测误差越大，预测值显著低于实际值，可以看出预测效果不好；上证 50 指数跟踪图可以看出跟踪效果不佳，与基于逐步回归的指数追踪图十分相似，两者无太大别。

5. 弹性约束估计建立上证 50 指数与成分股的回归方程

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。

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：

$$\tilde{\beta}_j = \arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right) \tag{7}$$

等价于找到使得下式

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \text{满足}(1-\lambda)\sum_{j=1}^{p}\left|\beta_j\right| + \lambda\sum_{j=1}^{p}\beta_j^2 \le t \tag{8}$$

达到最小的 $\beta_j, j = 1,2,\cdots,p$。

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。

用弹性约束估计建立上证 50 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 λ 值，$\lambda_{\min} = 0.5602381$，$\lambda$ 值的选择如图 3 所示。

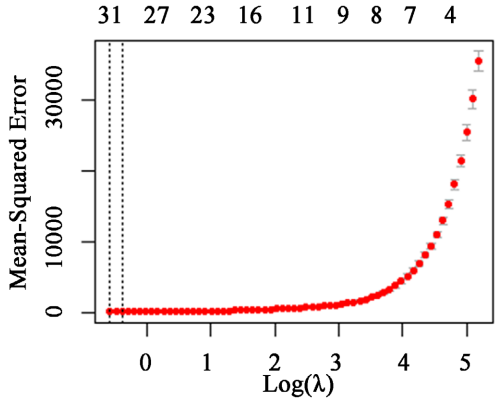

Figure 3. λ Selection graph
图 3. λ 选择图

图 3 中横轴是对数 λ 值，纵轴是均方误差。红点代表均方误差和上下一倍标准差，均方误差越小模型越好；上方数量表明模型仍存在的自变量个数(不一定是单调递减)。第一条虚线处表明均方误差最小值；第二个虚线标出最低点的一倍标准差的位置，表示牺牲一倍标准差的情况下可以得到的最简单的模型。按此参数值，保留变量个数是 28 个，残差平方和为 8105.894，对应的残差图如下图 4 所示。

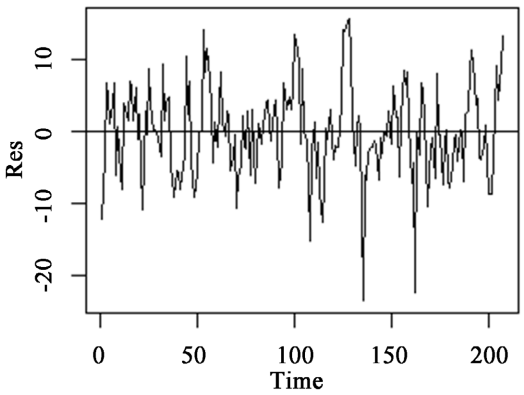

Figure 4. Residual diagram
图 4. 残差图

作为基金公司，需要用最少的变量达到对指数的跟踪，从而实现股票与股指期货的对冲，达到保值目的，这时太多的股票，全部持有几乎不可能，有必要进一步进行变量选择。

而本例通过 CV 准则得到的最优解仅仅包含 28 只成分股，已经能够满足持股和利用股指期货实现风险对冲的目的。根据 glmnet 程序此时给出的系数估计，得到回归模型如下所示，对应的基于弹性约束估计所选择的上证50指数上证(000016.SH)的28只成分股回归模型如下所示，表2为其对应的28只成分股。

$$\begin{aligned}\hat{Y}_3 = {} & 480 + 13.63X_2 + 5.71X_3 + 0.82X_4 + 6.24X_5 + 1.95X_7 + 3.55X_{10} + 1.26X_{11} + 2.61X_{14} + 0.27X_{15} \\ & + 0.92X_{16} + 0.17X_{18} + 4.97X_{19} + 0.35X_{21} + 6.55X_{22} + 0.41X_{23} + 7.72X_{26} + 0.18X_{27} + 2.40X_{28} \\ & + 2.92X_{30} + 5.78X_{33} + 0.91X_{41} + 0.10X_{42} + 12.53X_{43} + 0.85X_{44} + 0.20X_{46} + 0.30X_{47} + 0.19X_{48} + 0.28X_{50}\end{aligned} \tag{9}$$

Table 2. The 28 constituent stocks of the selected SSE 50 Index SSE (000016.SH) are estimated based on elastic constraints
表 2. 基于弹性约束估计所选择的上证 50 指数上证(000016.SH)的 28 只成分股

变量	名称	变量	名称	变量	名称	变量	名称
X2	中国石化	X14	通威股份	X23	伊利股份	X42	中国中免
X3	中信证券	X15	贵州茅台	X26	三峡能源	X43	紫金矿业
X4	三一重工	X16	恒生电子	X27	隆基绿能	X44	中远海控
X5	招商银行	X18	用友网络	X28	中信建投	X46	药明康德
X7	上汽集团	X19	海尔智家	X30	兴业银行	X47	海天味业
X10	恒瑞医药	X21	山西汾酒	X33	中国平安	X48	韦尔股份
X11	万华化学	X22	海通证券	X41	中国石油	X50	兆易创新

这些成分股在 2021 年下半年至 2022 上半年左右了上证 50 指数的走势，根据弹性约束建立的模型，完全可以模拟出上证 50 指数的实际走势，如图 5 所示。

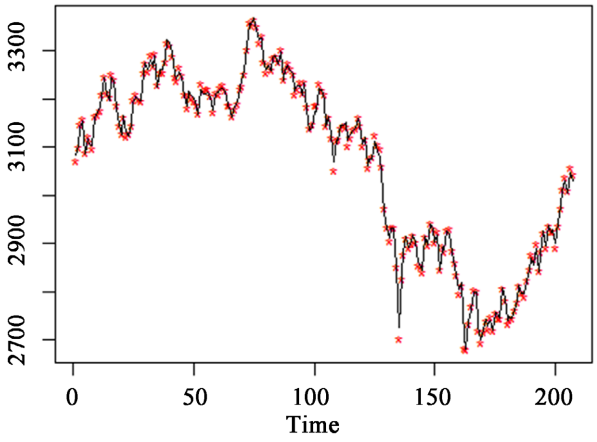

Figure 5. The tracking effect of the remaining 28 constituent stocks on the trend of SSE 50 index
图 5. 剩余 28 只成分股对上证 50 指数走势的跟踪效果

图 5 中星号是实际上证 50 指数每日的收盘价，实线是模型曲线。易见，实际走势和 28 只成分股的跟踪走势基本重合，基本跟踪到了指数的运行趋势，与指数运行合拍，对于以上证 50 为标的股指期货空单有较好的对冲效果。此外，还本文还采用 LASSO 交叉验证法。

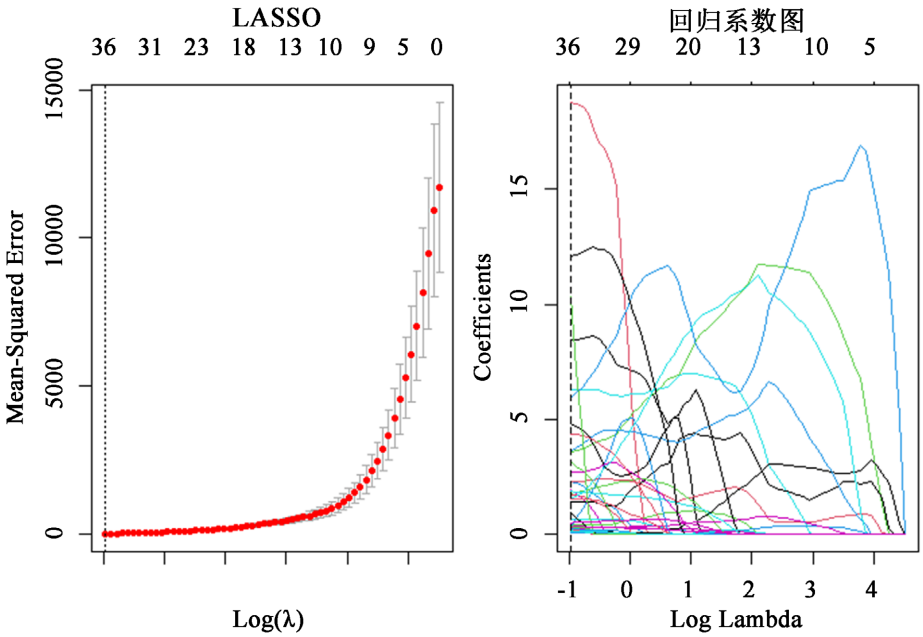

Figure 6. λ selection plot and regression coefficient plot
图 6. λ 选择图以及回归系数图

从图 6 中可知，当 $\lambda_{\min}=0.3795037$ 时，保留 36 个变量。其中 $\left(X_8,X_9,X_{12},X_{24},X_{29},X_{31},X_{32},X_{37},X_{38},X_{39},X_{40},X_{41},X_{45},X_{49}\right)$ 变量系数被压缩为 0，从而未被选入，相当于选择了 36 个变量。

$$\begin{aligned}\hat{Y}_4 &= 8.46X_1+18.78X_2+3.53X_3+2.30X_4+6.27X_5+1.93X_6+1.41X_7+2.29X_{10}+0.40X_{11}\\ &+0.05X_{13}+0.65X_{14}+0.27X_{15}+0.19X_{16}+0.25X_{17}+1.76X_{18}+3.56X_{19}+0.15X_{20}+0.23X_{21}\\ &+4.80X_{22}+1.72X_{23}+3.13X_{25}+0.31X_{26}+1.85X_{27}+2.67X_{28}+0.94X_{30}+4.34X_{33}+10.32X_{34}\\ &+5.92X_{35}+1.49X_{36}+0.35X_{42}+12.08X_{43}+1.50X_{44}+0.33X_{46}+0.21X_{47}+0.11X_{48}+0.49X_{50}\end{aligned} \tag{10}$$

在 $\hat{Y}_4$ 模型下对应的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图 7 如下。

对图 7 进行分析：从拟合残差图可知，该拟合残差在 0 上下波动，无明显趋势效应；从预测残差图显示，不是该残差白噪声序列；上证 50 指数预测图可以看出随着时间的增加，预测误差越大，预测值显著低于实际值，可以看出预测效果不好。

6. 岭回归指数追踪模型

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。

岭迹是指将 λ 从 0 增加到正无穷的过程得到的中每个分量的变化曲线。岭迹法选择 λ 值的一般原则是：1) 各回归系数的岭估计基本稳定；2) 用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理；3) 回归系数没有不合乎经济意义的绝对值；4) 残差平方和增大不多。

与 LASSO 相比，岭回归得到的模型一直都是 50 个变量，因此岭回归没有变量筛选的功能。运用 LASSO 进行变量选择，然后再通过最小二乘回归、岭回归、刘回归等对筛选出来的变量进行回归分析，

从而对指数进行跟踪得到模型。

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上证 50 指数的追踪效果最佳，某种方法下的上证 50 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。在下表 3 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。

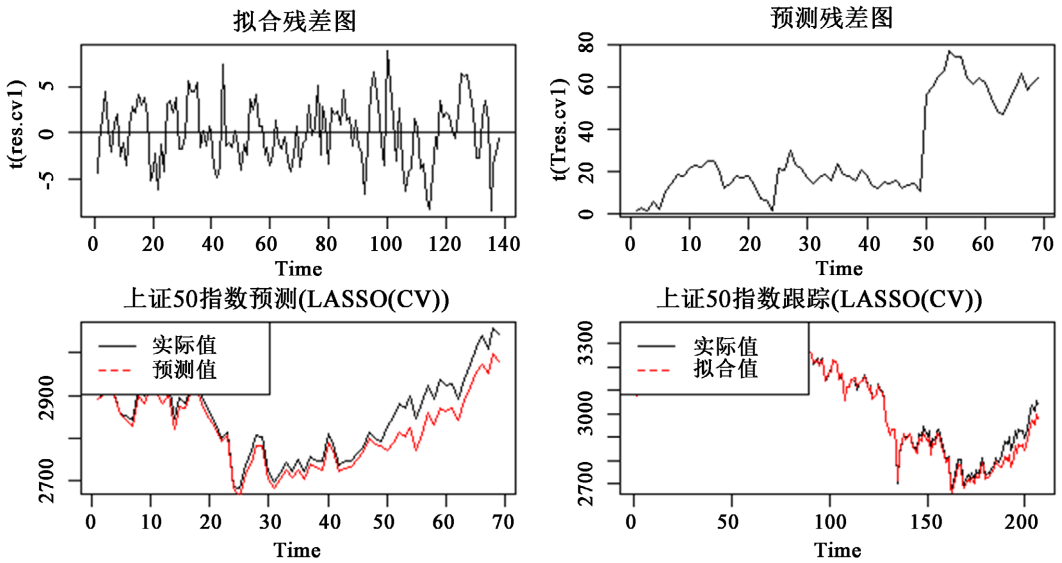

Figure 7. $\hat{Y}_4$ model of fitting residual graph, forecast residual graph, SSE 50 index forecast graph, SSE 50 index tracking graph

图 7. $\hat{Y}_4$ 模型的拟合残差图、预测残差图、上证 50 指数预测图、上证 50 指数跟踪图

Table 3. C_p comparison of index tracking effects of different methods under the criterion

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(43 个变量)	201.4673	2.14327	1.212668	105399.2	4215.967	27.51703
LASSO (49 个变量)	194.9932	2.215832	1.193025	91512.74	4816.46	26.17163
岭回归	200.1542	2.003084	1.197432	96531.38	5080.59	27.61190
LASSO + olse	831.9254	9.347477	2.464234	44111.98	2205.599	20.637
LASSO + 岭回归 (lambda = 0.03695123)	832.718	41.6359	2.465407	44656.91	2232.845	20.60334
LASSO + LIU (olse)	958.2969	10.76738	2.644781	39096.67	1954.833	18.81555

从表 3 中可知：在 C_p准则下，LASSO 保留了 49 个变量(成分股)，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，进一步运用刘估计进行回归，得到较好的外预测效果。为此，我们绘制出 Lasso + LIU (olse)方法下的指数跟踪图，如图 8 所示。

LASSO + LIU (olse)的回归方程如下：

$$\begin{aligned}\hat{Y}_5 &= 3.98X_1 + 0.18X_2 + 3.72X_3 + 2.71X_4 + 6.36X_5 + 3.02X_6 + 0.77X_7 - 0.56X_8 + 0.36X_9 + 3.36X_{10} + 0.99X_{11} \\ &+ 0.85X_{12} - 0.01X_{13} + 2.26X_{14} + 0.28X_{15} + 1.77X_{16} + 0.86X_{17} + 0.78X_{18} + 0.95X_{19} + 0.24X_{20} \\ &+ 0.29X_{21} + 2.82X_{22} + 0.46X_{23} + 4.29X_{24} - 3.63X_{25} + 1.70X_{26} + 1.40X_{27} + 0.70X_{28} \\ &+ 3.15X_{29} + 3.77X_{30} + 70.51X_{31} + 4.46X_{32} + 27.95X_{33} + 2.29X_{34} + 1.41X_{35} - 0.05X_{36} \\ &+ 8.77X_{37} + 0.30X_{38} + 0.41X_{39} + 4.78X_{40} + 10.33X_{41} + 0.66X_{42} - 0.15X_{43} + 0.64X_{44} \\ &+ 0.25X_{45} + 0.19X_{46} + 0.28X_{47} + 0.30X_{48}\end{aligned} \tag{11}$$

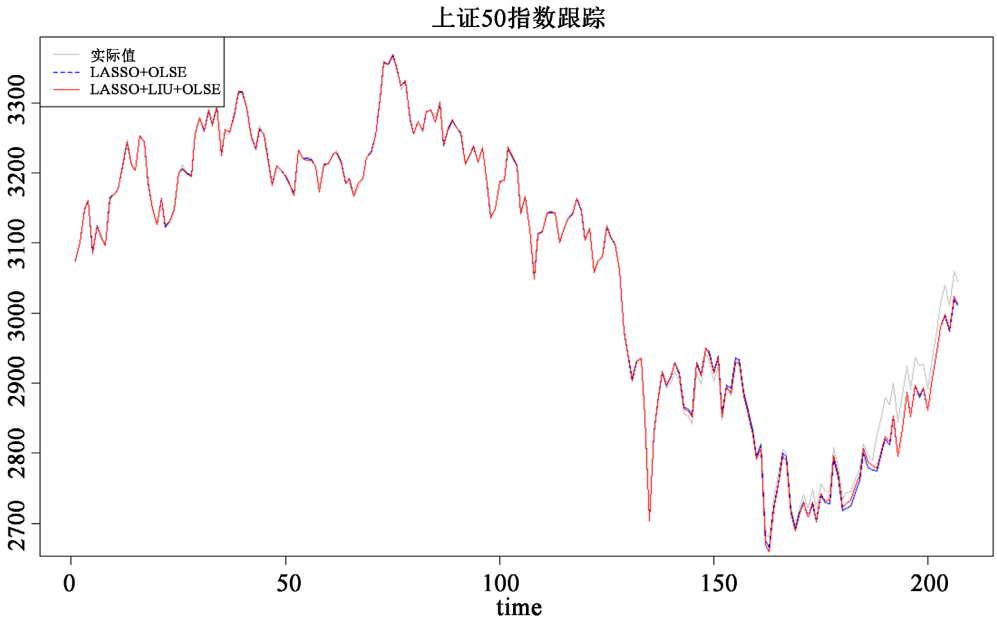

Figure 8. SSE 50 index tracking graph
图 8. 上证 50 指数跟踪图

Table 4. CV comparison of index tracking effects of different methods under the criterion
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(43 个变量)	201.4673	2.14327	1.212668	105,399.2	4215.967	27.51703
LASSO (36 个变量)	1579.216	15.6358	3.395162	93,333.86	2828.299	22.48379
岭回归(lambda = 9.184476)	4407.568	50.6617	5.672039	104,019.4	5778.857	16.4305
LASSO + olse	341.4063	3.38026	1.578612	137,194.7	4287.333	28.03984
LASSO + 岭回归 (lambda = 0.009873543)	341.442	10.67006	1.578695	137,131.8	4285.368	28.00536
LASSO + LIU (olse)	372.8135	3.691223	1.649626	144,740.5	4523.14	28.06844

从表 4 可以看出：相比于 C_p 准则，CV 准则下的 LASSO 只保留了 36 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，岭估计外预测效果也是较好的。为此，我们绘制出 LASSO + 岭回归方法下的指数跟踪图，如图 9 所示。

LASSO + 岭回归的回归方程：

$$\begin{aligned}\hat{Y}_6 &= 6.94X_1 + 18.88X_2 + 2.84X_3 + 2.68X_4 + 5.77X_5 + 3.82X_6 + 0.93X_7 + 2.50X_8 + 0.59X_9 + 0.08X_{10} \\ &+ 0.93X_{11} + 0.27X_{12} + 0.42X_{13} + 0.98X_{14} + 1.05X_{15} + 1.43X_{16} + 0.44X_{17} + 0.12X_{18} + 6.62X_{19} + 1.76X_{20} \\ &+ 4.78X_{21} - 4.31X_{22} + 2.00X_{23} + 2.86X_{24} + 4.33X_{25} + 4.55X_{26} + 52.71X_{27} + 3.49X_{28} - 0.36X_{29} + 0.48X_{30} \\ &+ 11.06X_{31} + 1.38X_{32} + 0.70X_{33} + 0.33X_{34} + 0.26X_{35} + 0.35X_{36}\end{aligned} \tag{12}$$

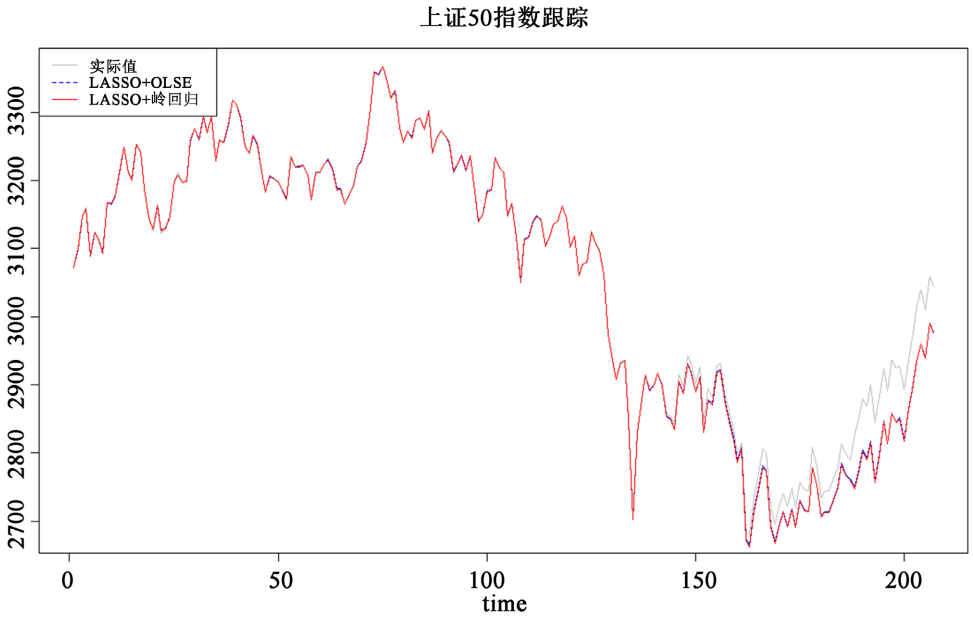

Figure 9. SSE 50 index tracking graph
图 9. 上证 50 指数跟踪图

7. 分位数回归指数追踪模型

在线性回归模型中，需要假设 X 和 Y 服从二元正态分布为前提，但是如果 X 和 Y 不服从二元正态分布，那么我们用来描述 $F\left(Y|X\right)$ 的分布就不仅仅只是条件期望和条件方差了。最方便的一个方法就是用分位数来描述 $F\left(Y|X\right)$ 的分布。

首先我们先给出在给定 X 下的 Y 的条件分位数的表达式，当然，此时的 X 和 Y 是服从任意联合分布的。假设 X 和 Y 仍然是通过简单线性模型相关.而误差项 ε 的分量 ε_i 是独立同分布的。但是此时需引入一个特定的误差分布函数，记作 F_ε。现在考虑条件分位数的线性回归模型，同时，仍假设 ε 的期望为 0，但是它的分位数不为 0。因此，当我们用分位数回归来代替线性回归模型时，误差项不会消失。

令 $q \in (0,1)$，并用 $F_\varepsilon^{-1}(q)$ 代表 q 分位点的误差。同时因变量的条件 q 分位点通过 $F\left(Y|X\right)=q$ 得到 $Y = F^{-1}\left(q|X\right)$，那么条件 q 分位点的线性回归模型为：

$$F^{-1}\left(q|X\right) = \alpha + \beta X + F_\varepsilon^{-1}(q) \tag{13}$$

而该公式就是一个分位数回归模型。

在分位数回归中还是通过一个散点图来画分位数回归线，分位数回归与普通线性回归的区别在于线性回归线穿过“平均或重心”的点，而分位数回归线则通过分位点。

由于作为基金公司，需要用最少的变量达到对指数的跟踪，从而实现股票与股指期货的对冲，达到保值目的，对比两个表格，发现基于方法(CV 准则)选择 LASSO (36 个变量)与基于方法(Cp 准则)选择 LASSO (49 个变量)的平均残差平方和(RMS)和残差标准差(SD)相差并不是很大，但是选择的变量个数相差了 13 个之多。为此，下面我将采用基于方法(CV 准则)选择 LASSO (36 个变量)进行。

当分位数为 $q=0.05$ 时获得经验回归方程如下：

$$\begin{aligned}\hat{Y}_7 &= -4.65+5.55X_1+17.27X_2+2.30X_3+2.96X_4+5.83X_5+3.78X_6+0.48X_7+2.49X_{10}+0.69X_{11}\\ &+0.09X_{13}+0.85X_{14}+0.28X_{15}+0.57X_{16}+1.04X_{17}+0.92X_{18}+0.48X_{19}+0.42X_{20}+0.14X_{21}+8.88X_{22}\\ &+2.18X_{23}+4.64X_{25}-5.54X_{26}+2.06X_{27}+2.73X_{28}+5.50X_{30}+4.91X_{33}+43.35X_{34}+4.26X_{35}-0.99X_{36}\\ &+0.45X_{42}+9.35X_{43}+2.03X_{44}+0.74X_{46}+0.25X_{47}+0.24X_{48}+0.37X_{50}\end{aligned} \tag{14}$$

当分位数为 $q=0.25$ 时获得经验回归方程如下：

$$\begin{aligned}\hat{Y}_8 &= -0.85+6.63X_1+15.20X_2+2.52X_3+2.46X_4+6.04X_5+3.69X_6+1.58X_7+2.57X_{10}+0.67X_{11}+0.10X_{13}\\ &+0.88X_{14}+0.28X_{15}+0.19X_{16}+1.09X_{17}+1.08X_{18}+1.39X_{19}+0.39X_{20}+0.09X_{21}+8.52X_{22}+1.52X_{23}\\ &+1.91X_{27}+2.98X_{28}+4.47X_{30}+4.57X_{33}+46.88X_{34}+5.04X_{25}-4.09X_{26}+3.94X_{35}-0.72X_{36}+0.43X_{42}\\ &+11.25X_{43}+1.63X_{44}+0.72X_{46}+0.37X_{47}+0.19X_{48}+0.42X_{50}\end{aligned} \tag{15}$$

当分位数为 $q=0.5$ 时获得经验回归方程如下：

$$\begin{aligned}\hat{Y}_9 &= 11.99+4.87X_1+17.22X_2+3.02X_3+2.35X_4+6.09X_5+3.81X_6+1.24X_7+2.34X_{10}+0.66X_{11}\\ &+0.08X_{13}+0.84X_{14}+0.27X_{15}+0.45X_{16}+1.34X_{17}+0.99X_{18}+1.85X_{19}+0.49X_{20}+0.12X_{21}+7.43X_{22}\\ &+1.41X_{23}+5.71X_{25}-5.41X_{26}+2.02X_{27}+2.80X_{28}+3.83X_{30}+4.47X_{33}+45.34X_{34}+2.82X_{35}+0.15X_{36}\\ &+0.48X_{42}+10.92X_{43}+1.10X_{44}+0.75X_{46}+0.38X_{47}+0.23X_{48}+0.33X_{50}\end{aligned} \tag{16}$$

当分位数为 $q=0.75$ 时获得经验回归方程如下：

$$\begin{aligned}\hat{Y}_{10} &= 22.86+9.35X_1+16.49X_2+3.31X_3+2.28X_4+5.80X_5+4.06X_6+0.93X_7+2.46X_{10}+0.58X_{11}+0.08X_{13}\\ &+0.91X_{14}+0.26X_{15}+0.48X_{16}+1.25X_{17}+1.16X_{18}+1.59X_{19}+0.43X_{20}+0.16X_{21}+7.78X_{22}+1.47X_{23}\\ &+5.07X_{25}-3.46X_{26}+1.95X_{27}+2.36X_{28}+4.05X_{30}+4.65X_{33}+46.68X_{34}+2.78X_{35}-0.41X_{36}+0.50X_{42}\\ &+11.02X_{43}+1.06X_{44}+0.72X_{46}+0.37X_{47}+0.28X_{48}+0.32X_{50}\end{aligned} \tag{17}$$

当分位数为 $q=0.95$ 时获得经验回归方程如下：

$$\begin{aligned}\hat{Y}_{11} &= 4.79+8.65X_1+24.06X_2+1.59X_3+2.01X_4+5.39X_5+3.68X_6+0.72X_7+2.50X_{10}+0.68X_{11}\\ &+0.09X_{13}+1.19X_{14}+0.26X_{15}+0.45X_{16}+1.13X_{17}+1.46X_{18}+1.23X_{19}+0.50X_{20}+0.09X_{21}+5.61X_{22}\\ &+1.80X_{23}+4.23X_{25}-4.28X_{26}+1.91X_{27}+3.67X_{28}+5.28X_{30}+4.32X_{33}+57.69X_{34}+2.79X_{35}+0.10X_{36}\\ &+0.52X_{42}+9.34X_{43}+0.81X_{44}+0.68X_{46}+0.34X_{47}+0.32X_{48}+0.32X_{50}\end{aligned} \tag{18}$$

并绘制了在不同分位数下的上证 50 指数跟踪图，如下图 10 所示。

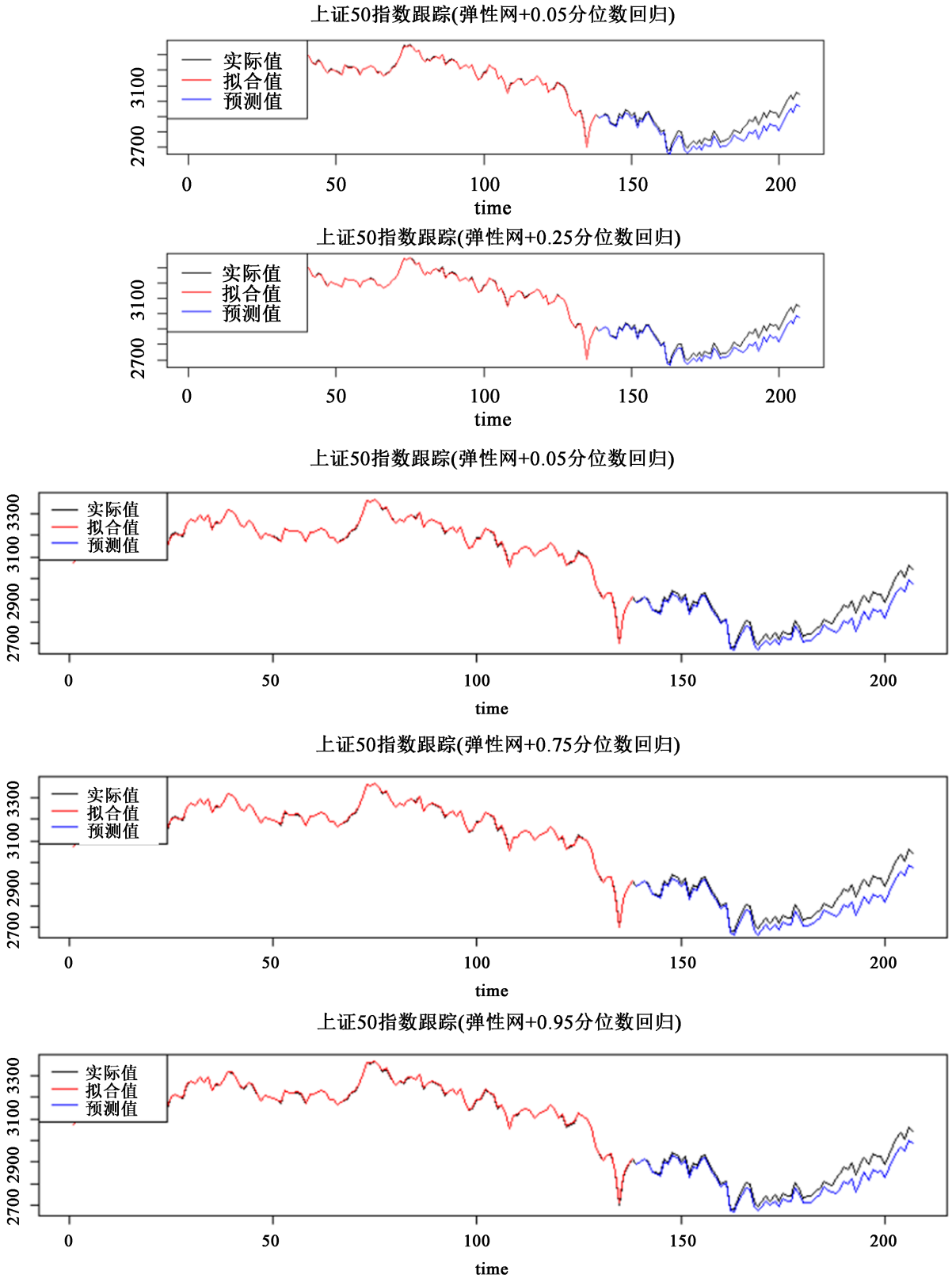

Figure 10. SSE 50 index tracking graph
图 10. 上证 50 指数跟踪图

通过上图 10 似乎无法判别基于哪个分位数的上证 50 指数跟踪效果更好，为此，我们通过计算在不同分位数下的指数跟踪的残差平方和、平均残差平方和、残差标准差来进行评价。

Table 5. Comparison of index tracking effect under different quantiles
表 5. 不同分位数下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(43 个变量)	201.4673	2.14327	1.212668	105,399.2	4215.967	27.51703
LASSO (36 个变量)	1579.216	15.6358	3.395162	93,333.86	2828.299	22.48379
岭回归(lambda = 9.184476)	4407.568	50.6617	5.672039	104,019.4	5778.857	16.4305
LASSO + 0.05 分位数回归	1101.100371	10.79510167	3.285590	183,661.3	5565.493	29.62569
LASSO + 0.25 分位数回归	591.95227	5.803454	2.4090358	138,480.2	4196.37	28.57266
LASSO + 0.5 分位数回归	417.6217	4.094331	2.023445	133,381.1	4041.852	28.30217
LASSO + 0.75 分位数回归	598.022632	5.86296698	2.421356	150,362.6	4556.443	28.75382
LASSO + 0.95 分位数回归	1105.00744	10.833406	3.2914140	112,017.4	3394.467	25.24244

从表 5 中可看出，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，0.5 分位数回归外预测效果也是最好的，0.05 分位数回归的效果最差。此外，我们还将基于上述所有方法所得模型的评价标准总结于表 6 中。

Table 6. Summary table of index tracking effect comparison under different models
表 6. 不同模型下的指数追踪效果对比汇总表

方法(C_p 准则)	训练集			测试集		
	SSE	RMS	SD	SSE	RMS	SD
逐步回归(43 个变量)	201.4673	2.14327	1.212668	105,399.2	4215.967	27.51703
LASSO (49 个变量)	194.9932	2.215832	1.193025	91,512.74	4816.46	26.17163
岭回归	200.1542	2.003084	1.197432	96,531.38	5080.59	27.61190
LASSO + olse	831.9254	9.347477	2.464234	44,111.98	2205.599	20.637
LASSO + 岭回归 (lambda = 0.03695123)	832.718	41.6359	2.465407	44,656.91	2232.845	20.60334
LASSO + LIU (olse)	958.2969	10.76738	2.644781	39,096.67	1954.833	18.81555
方法(CV 准则)	训练集			测试集		
	SSE	RMS	SD	SSE	RMS	SD
逐步回归(43 个变量)	201.4673	2.14327	1.212668	105,399.2	4215.967	27.51703
LASSO (36 个变量)	1579.216	15.6358	3.395162	93,333.86	2828.299	22.48379
岭回归 (lambda = 9.184476)	4407.568	50.6617	5.672039	104,019.4	5778.857	16.4305
LASSO + olse	341.4063	3.38026	1.578612	137,194.7	4287.333	28.03984
LASSO + 岭回归 (lambda = 0.009873543)	341.442	10.67006	1.578695	137,131.8	4285.368	28.00536

Continued

LASSO + LIU (olse)	372.8135	3.691223	1.649626	144,740.5	4523.14	28.06844
LASSO + 0.05 分位数回归	1101.100371	10.79510167	3.285590	183,661.3	5565.493	29.62569
LASSO + 0.25 分位数回归	591.95227	5.803454	2.4090358	138,480.2	4196.37	28.57266
LASSO + 0.5 分位数回归	417.6217	4.094331	2.023445	133,381.1	4041.852	28.30217
LASSO + 0.75 分位数回归	598.022632	5.86296698	2.421356	150,362.6	4556.443	28.75382
LASSO + 0.95 分位数回归	1105.00744	10.833406	3.2914140	112,017.4	3394.467	25.24244

从表 6 中可以看出，1) 在 Cp 准则下，LASSO 保留了 49 个变量(成分股)，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，进一步运用刘估计进行回归，得到较好的外预测效果。2) 在 CV 准则，LASSO 只保留了 36 个变量，外预测效果明显都次于 Cp 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，岭估计外预测效果也是较好的。3) 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，0.5 分位数回归外预测效果也是最好的，0.05 分位数回归的效果最差。4) 在不同的选股法下值的改变对模型的影响不同。

8. 总结

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，本文通过 Cp 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、岭回归、两步回归模型和分位数回归模型等。通过研究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。

参考文献

[1] Roll, R. (1992) A Mean/Variance Analysis of Tracking Error. *Journal of Portfolio Management*, **18**, 13-22. https://doi.org/10.3905/jpm.1992.701922

[2] Alexander, G.J. and Baptista, A.M. (2010) Active Portfolio Management with Benchmarking: A Frontier Based on Alpha. *Journal of Banking & Finance*, **34**, 2185-2197. https://doi.org/10.1016/j.jbankfin.2010.02.005

[3] Gotoh, J.Y. and Takeda, A. (2011) On the Role of Norm Constraints in Portfolio Selection. *Computational Management Science*, **8**, 323.

[4] 梁斌, 陈敏, 缪柏其, 等. 基于 LARS-Lasso 的指数跟踪及其在股指期货套利策略中的应用[J]. 数理统计与管理, 2011, 30(6): 10.

[5] Liu, R.Z. and Zhou, Y. (2015) The Portfolio of Index Tracing and Index Predictability under Multi-Information—Based on Adaptive LASSO and ARIMA-ANN Method. *Systems Engineering*, **33**, 1-7.

[6] 苏治, 方彤, 秦磊. 一种基于规则化方法的最优稀疏指数追踪模型设计[J]. 数量经济技术经济研究, 2016, 33(4): 145-160.

[7] 胡梦婷. 基于回归模型的指数追踪问题及实证分析[D]: [硕士学位论文]. 昆明: 云南财经大学, 2017.

[8] 马景义, 单璐琪, 方彤. 一种增强型指数追踪模型设计及应用[J]. 数量经济技术经济研究, 2017, 34(5): 107-121.

[9] 彭胜银. 基于 Lasso 分位数的非负两阶段方法及在标普 500 指数追踪的应用[D]: [硕士学位论文]. 重庆: 重庆大学, 2019.