Operations Research and Fuzziology 运筹与模糊学, 2023, 13(4), 3299-3313
Published Online August 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.134333

Hans汉斯

基于粒子群优化算法的股票指数追踪及投资分析

陈　鑫

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年6月16日；录用日期：2023年8月4日；发布日期：2023年8月10日

摘　要

本文首先将二进制粒子群特征选择算法结合岭估计、Lasso估计和最小二乘估计利用成分股构建了三个关于上证50指数的指数追踪模型，三个模型都取得了非常优秀的追踪效果，其中无论是从追踪效果的角度还是投资角度出发，基于二进制粒子群特征选择算法结合最小二乘估计所得到的指数追踪模型都是最佳模型。然后利用三个指数追踪模型提取的成分股计算对上证50趋势的影响排名，其中贵州茅台的影响排名最高。最后在一定条件下进行投资行为的模拟和收益分析，发现基于三个指数追踪模型所提取的优质成分股进行投资可以获得比直接投资上证50指数更高的收益，达到了降低投资成本和投资风险的目的，为证50指数的投资提供科学合理的建议。

关键词

指数追踪，二进制粒子群，特征选择，岭估计，Lasso估计，最小二乘估计

Stock Index Tracking and Investment Analysis Based on Particle Swarm Optimization Algorithm

Xin Chen

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jun. 16th, 2023; accepted: Aug. 4th, 2023; published: Aug. 10th, 2023

Abstract

In this paper, we first combine the binary particle swarm feature selection algorithm with ridge

文章引用: 陈鑫. 基于粒子群优化算法的股票指数追踪及投资分析[J]. 运筹与模糊学, 2023, 13(4): 3299-3313.
DOI: 10.12677/orf.2023.134333

estimation, Lasso estimation and least square estimation to construct three index tracking models on SSE 50 Index using constituent stocks. The three models have achieved excellent tracking effects. The exponential tracking model based on binary particle swarm optimization feature selection algorithm and least squares estimation is the optimal model, whether from the perspective of tracking effect or investment. Then we use the constituent stocks extracted from the three index tracking models to calculate the impact ranking on the Shanghai Stock Exchange 50 trend, of which Kweichow Moutai ranks the highest, Finally, under certain conditions, the simulation and return analysis of investment behavior are carried out, and it is found that the investment of high-quality constituent stocks extracted based on the three index tracking models can obtain higher returns than direct investment in the SSE 50 Index, achieving the purpose of reducing investment costs and investment risks, and providing scientific and reasonable suggestions for the investment of the 50 Index.

Keywords

Exponential Tracking, Binary Particle Swarm, Feature Selection, Ridge Estimation, Lasso Estimation, Least Squares Estimation

1. 引言

近年来随着基金管理的日渐成熟，市场管理制度的不断完善，指数追踪[1]备受国内外投资者追捧。指数追踪的目的是从目标指数所包含的成分股中选择一组股票复制和跟踪目标指数的走势，从而获取与市场目标指数表现相近的收益，使得持有的股票与股指期货空单形成完全对冲以达到保值的效果。对于指数型基金管理者来说，常面临如何选取股票的问题，常见的方法有两种，分别是完全复制方法[2]和不完全复制方法[3]。若采用完全复制方法，则需购买目标指数的所有成分股票，根据每种成分股在目标指数中所占的权重来构建追踪组合，实现零误差复制目标指数收益率，但这种方法会产生较大的交易成本和高额的市场管理费用，特别是成分股较多的指数，如上证 50 指数包含了 50 只成分股，使用完全复制方法不符合实际；如果采用不完全复制方法，那么对上证 50 指数的优质成分股的提取及权重优化的分析与研究具有非常大的意义。本文将利用启发式粒子群优化算法[4]结合传统回归建模方法以最小化追踪组合走势和目标指数走势之间的误差获取优质成分股，以获得与指数差不多的收益率。虽然非完全复制方法存在一定的追踪误差，但投入的成本低，降低了投资风险，为较短期内投资者的投资选择提供合理可行的建议。

2. 数据介绍

本文研究数据来源于国内某证券交易网站，选取 2021 年 6 月 10 日至 2023 年 5 月 31 日上证 50 指数及其成分股日收盘价共 479 条数据作为研究对象。

上证 50 指数是上海证券交易所编制的一种综合指数，是从上市的所有股票中选取最具代表性的 50 只股票作为计算对象，并按成分股的调整股本数为权重进行加权计算得出的加权股价指数，其综合反映了上海证券市场最具市场影响力的一批龙头企业的整体状况。

本文研究的上证 50 指数成分股所属公司如表 1 所示：

Table 1. List of 50 constituent stocks on the Shanghai Stock Exchange
表 1. 上证 50 成分股列表

变量	公司	变量	公司	变量	公司
x_1	天合光能	x_{18}	中国太保	x_{35}	贵州茅台
x_2	兆易创新	x_{19}	工商银行	x_{36}	通威股份
x_3	华友钴业	x_{20}	中国平安	x_{37}	片仔癀
x_4	韦尔股份	x_{21}	农业银行	x_{38}	国电南瑞
x_5	海天味业	x_{22}	陕西煤业	x_{39}	恒力石化
x_6	合盛硅业	x_{23}	兴业银行	x_{40}	万华化学
x_7	药明康德	x_{24}	中国神华	x_{41}	恒瑞医药
x_8	中金公司	x_{25}	中信建投	x_{42}	复星医药
x_9	中远海控	x_{26}	隆基绿能	x_{43}	北方稀土
x_{10}	紫金矿业	x_{27}	三峡能源	x_{44}	上汽集团
x_{11}	中国中免	x_{28}	长江电力	x_{45}	保利发展
x_{12}	中国石油	x_{29}	航发动力	x_{46}	招商银行
x_{13}	华泰证券	x_{30}	伊利股份	x_{47}	三一重工
x_{14}	中国电建	x_{31}	山西汾酒	x_{48}	中信证券
x_{15}	中国建筑	x_{32}	闻泰科技	x_{49}	中国石化
x_{16}	长城汽车	x_{33}	海尔智家	x_{50}	包钢股份
x_{17}	中国人寿	x_{34}	海螺水泥		

3. 研究目的及分析流程

3.1. 研究目的

以上证 50 指数各成分股日收盘价作为自变量，优化成分股权重并构建上证 50 指数日收盘价的指数追踪模型，剔除不显著的成分股，最终提取出对上证 50 指数波动具有重要影响的优质成分股及其影响排名，并对其进行投资模拟分析，为上证 50 指数的投资提供科学合理的建议，减少投资者的交易成本和降低交易风险。

3.2. 分析流程

具体分析流程如图 1 所示。

第一步：通过通信达金融终端获取本文所研究的数据集，上证 50 指数及其成分股日收盘价。

第二步：将数据集按照时间先后顺序进行训练集(70%)和测试集(30%)划分。

第三步：将二进制粒子群特征选择算法分别结合岭估计、Lasso 估计和最小二乘估计构建三个指数追踪模型并得到优质成分股。

第四步：对三个指数追踪模型的优质成分股进行影响分析并得到影响排名。

第五步：基于影响排名进行投资行为模拟并进行收益分析。

第六步：给出结论及投资建议。

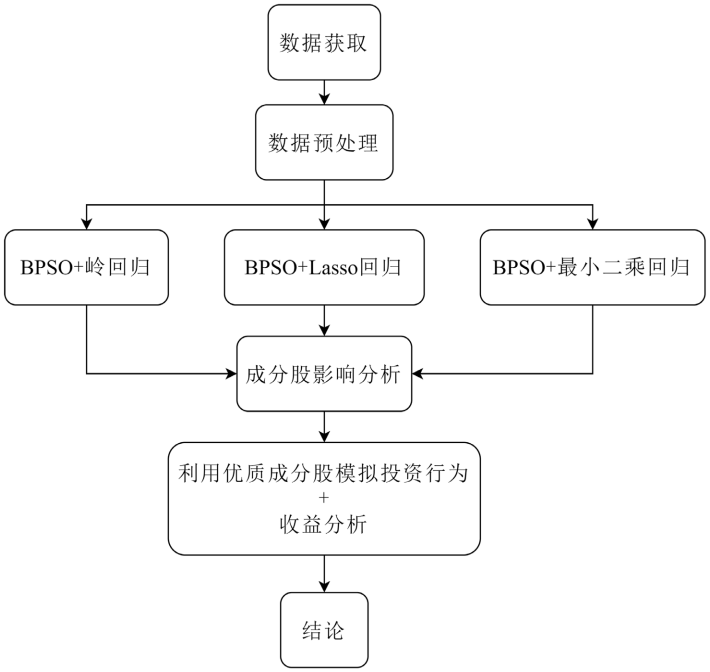

Figure 1. Analysis flow chart
图 1. 分析流程图

4. 研究方法

4.1. 多元线性模型

多元线性模型通常用来研究一个因变量依赖多个自变量的变化关系，如果二者的依赖关系可以用线性形式来刻画，则可以建立多元线性模型来进行分析与研究。

模型定义：

多元线性模型通常用来描述变量 y 与 x 之间的随机线性关系，即

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon \tag{1}$$

式中，$x_1,\cdots,x_p$ 是非随机的自变量；y 是随机的因变量；β_0 是常数项；$\beta_1,\cdots,\beta_p$ 是回归系数；ε 是随机误差项。

如果对 y 和 x 进行了 n 次观测，得到 n 组观测值 $y_i,x_{1i},\cdots,x_{pi}\left(i=1,2,\cdots,n\right)$，它们满足以下关系式

$$y_i=\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pn}+\varepsilon_i \tag{2}$$

引入矩阵记号，记

$$\boldsymbol{y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\ \boldsymbol{X}=\begin{bmatrix}1&x_{11}&\cdots&x_{p1}\\1&x_{12}&\cdots&x_{p2}\\\vdots&\vdots&\ddots&\vdots\\1&x_{1n}&\cdots&x_{pn}\end{bmatrix},\ \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{bmatrix},\ \boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$$

则，模型(2)可以下位如下形式

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{3}$$

式中，$\boldsymbol{y}$ 是观测向量；$\boldsymbol{X}$ 是已知的设计矩阵；$\boldsymbol{\beta}$ 是未知参数向量；$\boldsymbol{\varepsilon}$ 是随机误差向量。

如果模型(3)满足条件：1) $E(\boldsymbol{\varepsilon})=0$，2) $Var(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$，3) $x_1,\cdots,x_p$ 互不相关，则称模型(3)为普通线性回归模型。

在正态假定下，如果 $\boldsymbol{X}$ 是列满秩的，则普通线性回归方程(3)的参数 $\boldsymbol{\beta}$ 的最小二乘估计为：

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{4}$$

于是得到回归方程

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{5}$$

4.2. 岭回归

1970 年霍尔(Hoerl)和肯纳德(Kennard) [5]提出了岭估计。可以有效解决传统回归模型在设计矩阵病态或者变量之间存在多重共线性时不再适用的问题。

考虑线性模型 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$，样本数据为 $y_i,x_{1i},\cdots,x_{pi}\left(i=1,2,\cdots,n\right)$，误差向量满足 $E(\boldsymbol{\varepsilon})=0$，$Var(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$，假设这些样本是相互独立的，或者在给定 x_{ij} 的情况下 y_j 是独立的，同时也假设所有的 x_{ij} 都进行了标准化处理，即

$$\sum_{j=1}^{n}x_{ij}=0,\ \sum_{j=1}^{n}x_{ij}^2=1 \tag{6}$$

可以通过解决如下条件极值问题获得

$$\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)+k\boldsymbol{\beta}'\boldsymbol{\beta} \tag{7}$$

设 $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)'$，则由(7)，岭估计其实就是使下式达到最小的参数估计

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ji}\right)^2+k\sum_{j=1}^{p}\beta_j^2 \tag{8}$$

其中 k 是拉格朗日乘数(Lagrangian Multipliers)，岭估计有如下表达式

$$\hat{\boldsymbol{\beta}}(k)=\left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{9}$$

其中 $k\geq 0$ 是岭参数。岭估计的分量 $\hat{\beta}_i(k)$ 作为 k 的函数，通过对 k 值的选择，可以减少多重共线性的影响，取不同的 k 值，可以得到不同的回归系数估计，因此岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是一个关于 k 的估计类。当 k 在 $[0,+\infty)$ 变化时，在平面直角坐标系所描出的图形称为岭迹，选择 k 的岭迹法是：将 p 个分量 $\hat{\beta}_i(k)$ 的岭迹画在同一个图上，选择 k 使得各个分量的值大致稳定，并且兼顾回归系数没有不合理的符号，残差平方和上升不太多等。当基于以上方法选择一个 k 值之后便可以得到岭估计 $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)$，于是得到回归方程

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{10}$$

需要注意的是，当数据是标准化的时候有 $\bar{y}=0$，又有 β_0 解 $\hat{\beta}_0=\bar{y}=0$，所以回归方程不含常数项。

4.3. 绝对约束估计

1996 年蒂贝希拉尼(Tibashirani) [6]提出了一种可以具有变量选择功能的估计方法——绝对约束估计(Lasso: the least absolute shrinkage and selection operator)。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，其等价于求下式达到最小值

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ji}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{11}$$

其中 $\lambda\geq 0$，Lasso 的复杂程度由 λ，λ 越大对变量较多的线性模型的惩罚力度就越大，从而最终获得更加简洁的模型。一般可以通过 Cp 准则、AIC 准则、BIC 准则和 CV 交叉验证得到最优 λ 下的最优回归模型。

4.4. 粒子群优化算法(PSO: Particle Swarm Optimization)

J. Kennedy 和 R. Eberhart [7]等受到鸟类觅食的集群活动的启发而于 1995 年开发的一种演化计算技术.粒子群优化算法的思想是：假设空间中有一些粒子随机的分布在不同的空间位置上，每个粒子有两个属性——位置和速度。这两个属性的初始值都是随机的，位置代表了问题的解，而速度代表了每一次迭代中解的变化方向和快慢。在使用粒子群优化算法时，还要根据需要解决的问题设定合适的目标函数，根据目标函数可以计算出当前迭代为止每个粒子的最优解 $pbest_i$ 以及所有粒子的全局最优解 $gbest$。每个粒子维护一个自身最优点 $pbest_i$，整个粒子群维护一个全局最优点 $gbest$。然后进行迭代，迭代的过程可以描述为粒子按照自身最优 $pbest_i$ 的方向和全局最优 $gbest$ 的方向按一定比例来改变速度，继而改变本次迭代粒子到达的位置。公式如下所示：

速度公式：

$$v_i^{t+1}=w*v_i^t+c_1*rand1*\left(pbest_i-x_i^t\right)+c_2*rand2*\left(gbest-x_i^t\right) \tag{12}$$

其中 w、c_1 和 c_2 为权重参数，$rand$ 是来自[0, 1]区间的随机数，t 代表迭代次数，i 代表第 i 个粒子。

位置公式：

$$x_i^{t+1}=x_i^t+v_i^{t+1} \tag{13}$$

位置公式就是简单的匀速直线运动公式(一次迭代相当于经过一个时间单位)。

4.5. 二进制粒子群优化算法(BPSO: Binary Particle Swarm Optimization)

粒子群优化算法是一个解决连续空间问题的算法，而要想应用到特征选择问题上需要做一些细微的修改以使它能解决离散空间问题。J. Kenedy 和 R. Eberhart 在 1997 年开发出一个离散二进制版本简称(BPSO) [8]，改进方法是速度公式不变，增加了一个二进制公式 sigmoid 函数，重新定义了位置公式。如下：

$$x_{ij}=\begin{cases}1, & rand<Sigmoid\left(v_{ij}\right)\\ 0, & \text{otherwise}\end{cases},\quad Sigmoid\left(v_{ij}\right)=\frac{1}{1+e^{-v_{ij}}} \tag{14}$$

经过二进制化后，粒子的位置 x 为仅包含 0 和 1 元素的向量，就具有了特征选择功能。$gbest$ 中为 1 所对应的特征子集就是最终要选择的最优特征子集。

4.6. 模型评估指标

误差平方和(Sum of Squares due to Error, SSE)是在线性模型中衡量模型拟合程度的一个指标。模型的误差平方和越小表示模型拟合效果越好。其计算公式如下：

$$SSE=\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2 \tag{15}$$

平均绝对误差(Mean of Absolute Error, MAE)描述了真实值与预测值之间误差的均值，因而可以准确反映实际预测误差的大小。模型的平均绝对误差越小表示模型拟合效果越好。其计算公式如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{16}$$

平均相对误差(Mean of Relative Error，MRE)指相对误差的平均值，这个平均相对误差一般是用绝对值，平均相对误差更能反映预测的可信程度，相对误差一般以百分数的形式进行表示。模型的平均相对误差越小表示模型拟合效果越好。其计算公式如下：

$$\mathrm{MRE}=\frac{100}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i-y_i\right|}{y_i} \tag{17}$$

风险误差(Risk Error, RE)衡量了模型的预测稳定性。模型的风险误差越小表示模型预测效果更加稳定。其计算公式如下：

$$\mathrm{RE}=\sqrt{\frac{\sum_{i=1}^{n}\left(\left(\hat{y}_i-y_i\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)\right)^2}{n-1}} \tag{18}$$

5. 实证分析

5.1. 划分训练集和测试集

本文所研究股票数据属于时间序列类型数据，为了更好的对上证 50 指数构建回归模型进行指数追踪，需要将数据按照时间先后顺序进行训练集(70%)和测试集(30%)划分，其中 2021 年 6 月 10 日至 2022 年 10 月 28 日共 336 条数据作为训练集，2022 年 10 月 31 日至 2023 年 5 月 31 日共 143 条数据作为测试集。以训练集构建上证 50 指数追踪模型，利用测试集对指数模型进行质量评估。

5.2. 岭回归结合 BPSO

在构建岭回归方程之前需要确定参数 λ，λ 的确定现在已经有了很多相对比较成熟的方法。本文基于广义交叉验证 GCV [9]最小值来确定岭回归建模时所使用的 λ 值。

本文首先利用向后选择法逐步剔除不显著的自变量，剔除的原则是：每次剔除最不显著的自变量，即剔除检验统计量绝对值最小或者 P 值最大所对应的自变量，直至剩余的自变量对因变量都有显著的影响为止。然后基于最小化岭回归模型在测试集预测误差下，结合二进制粒子群优化算法对剩余成分股进行变量筛选，最终共剔除 24 只成分股。最后利用剩余成分股构建关于上证 50 指数的岭回归方程如下所示：

$$\begin{aligned}\hat{y}=&185.504-0.149x_1+0.684x_2+0.680x_3+0.795x_7+\cdots\\&+0.435x_{43}+8.600x_{44}-0.341x_{45}+5.980x_{48}+2.516x_{49}\end{aligned} \tag{19}$$

利用该回归方程可以得到基于 BPSO 特征选择之后的岭回归模型的模型检验结果如表 2 所示：

Table 2. Testing results of ridge regression model
表 2. 岭回归模型的检验结果

模型显著性检验 P 值	R-squared	Adjusted R-squared	残差平方和
< 2.2e −16	0.9952	0.9943	119742.5

模型在测试集上各评价指标值如表 3 所示：

Table 3. Evaluation results of ridge regression model
表 3. 岭回归模型的评估结果

SSE	MAE	MRE	风险误差
22544.6	10.1413	0.3826	12.4995

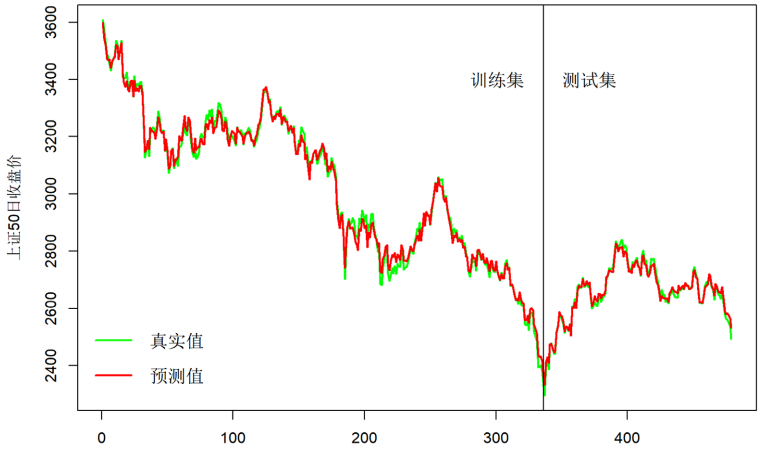

Figure 2. SSE 50 Index tracking chart of ridge regression model
图 2. 岭回归模型的上证 50 指数追踪图

从表 2 的模型检验结果可以发现，岭回归指数追踪模型通过了显著性检验同时解释了成分股和上证 50 指数之间绝大部分的线性关系。从表 3 的模型评估结果以及图 2 指数追踪图可以发现，基于二进制粒子群优化算法和岭估计所得到指数追踪模型是本文所构建的三个指数追踪模型中追踪能力和稳定性最差的，但就该模型而言其 MRE 指标为 0.3826，表明平均预测误差为真实值的 0.3826%，同样也展现了非常良好的指数追踪能力。

5.3. Lasso 回归结合 BPSO

在构建 Lasso 回归方程之前需要确定参数 λ，Lasso 回归方程复杂度由参数 λ 来控制，λ 越大对变量较多的线性模型的惩罚力度[10]就越大，从而最终获得一个更加简约的模型。λ 的确定现在已经有了很多相对比较成熟的方法。本文基于 CV 交叉验证法来确定 Lasso 回归建模时所使用的 λ 值，由此可获得自变量的回归系数估计值，对于回归系数为 0 的自变量认为其对于上证 50 指数日收盘价的影响不够显著，作出剔除处理。

本文首先利用 Lasso 估计自带的变量选择功能剔除回归系数估计为 0 的成分股。然后基于最小化 Lasso 回归模型在测试集预测误差下，结合二进制粒子群优化算法对剩余成分股进行变量筛选，最终共剔除 25 只成分股。最后利用剩余成分股构建关于上证 50 指数的 Lasso 回归方程如下所示：

$$\begin{aligned}\hat{y} = {} & 141.1074 + 0.6692x_1 + 0.6703x_2 + 0.3809x_3 + 2.0402x_5 + \cdots \\ & + 7.6016x_{44} + 2.2725x_{45} + 1.6886x_{47} + 10.5601x_{48} + 0.0058x_{50}\end{aligned} \tag{20}$$

利用该回归方程可以得到基于 BPSO 特征选择之后的 Lasso 回归模型的模型检验结果如表 4 所示：

Table 4. Testing results of lasso regression model
表 4. Lasso 回归模型的检验结果

模型显著性检验 P 值	R-squared	Adjusted R-squared	残差平方和
$< 2.2e-16$	0.9985	0.9982	34187.98

模型在测试集上各评价指标值如表 5 所示：

Table 5. Evaluation results of lasso regression model
表 5. Lasso 回归模型的评估结果

SSE	MAE	MRE	风险误差
8247.38	6.1209	0.2323	7.1286

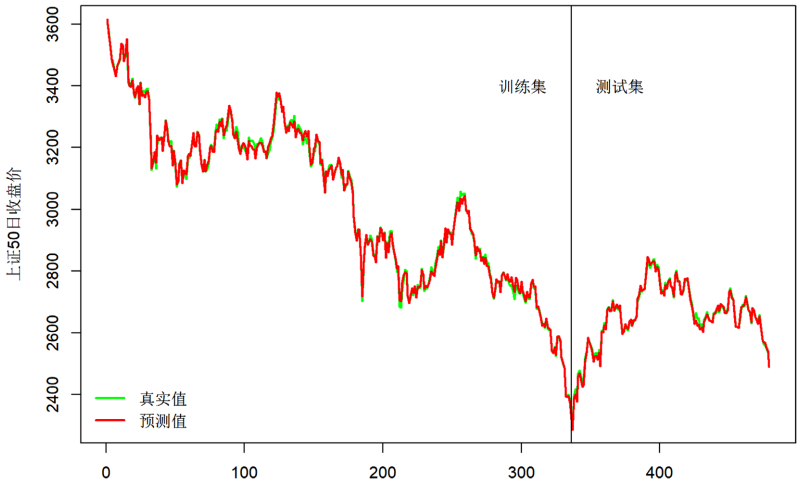

Figure 3. SSE 50 Index tracking chart of Lasso regression model
图 3. Lasso 回归模型的上证 50 指数追踪图

从表 4 的模型检验结果可以发现，Lasso 回归指数追踪模型通过了显著性检验同时解释了成分股和上证 50 指数之间绝大部分的线性关系。从表 5 的模型评估结果以及图 3 指数追踪图可以发现，基于二进制粒子群优化算法和 Lasso 估计所得到的指数追踪模型的指数追踪能力和稳定性在三个模型中都处于中间水平。

5.4. 最小二乘回归结合 BPSO

本文首先利用向后选择法逐步剔除不显著的自变量，剔除的原则是：每次剔除最不显著的自变量，即剔除检验统计量绝对值最小或者 P 值最大所对应的自变量，直至剩余的自变量对因变量都有显著的影响为止。然后基于最小化最小二乘回归模型在测试集预测误差下，结合二进制粒子群优化算法对剩余成分股进行变量筛选，最终共剔除 17 只成分股。最后利用剩余成分股构建关于上证 50 指数的最小二乘回归方程如下所示：

$$\begin{aligned}\hat{y} = {} & 136.7298 + 0.5166x_1 + 0.4954x_4 + 0.8881x_5 + 0.6115x_7 + \cdots \\ & + 4.4490x_{44} + 1.5898x_{45} + 4.2222x_{46} + 1.9406x_{47} + 0.1668x_{49}\end{aligned} \tag{21}$$

利用该回归方程可以得到基于 BPSO 特征选择之后的 Lasso 回归模型的模型检验结果如表 6 所示：

Table 6. Testing results of least squares regression model
表 6. 最小二乘回归模型的检验结果

模型显著性检验 P 值	R-squared	Adjusted R-squared	残差平方和
$< 2.2e-16$	0.9996	0.9995	9435.207

模型在测试集上各评价指标值如表 7 所示：

Table 7. Evaluation results of least squares regression model
表 7. 最小二乘回归模型的评估结果

SSE	MAE	MRE	风险误差
3091.778	3.7104	0.1405	4.4419

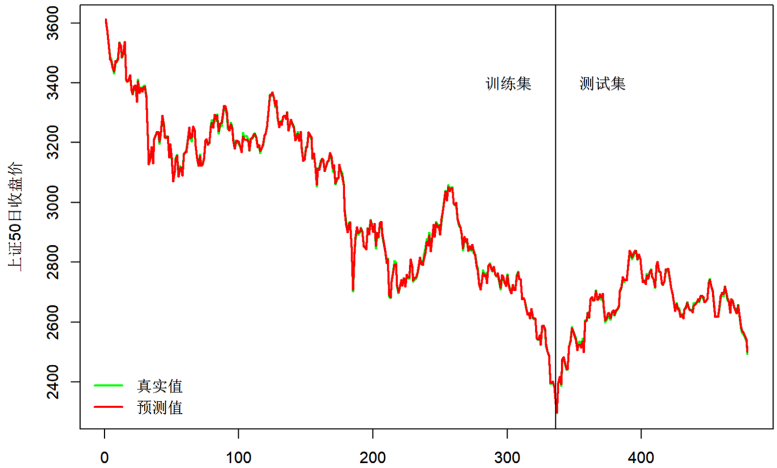

Figure 4. SSE 50 Index tracking chart of least squares regression model
图 4. 最小二乘回归模型的上证 50 指数追踪图

从表 6 的模型检验结果可以发现，最小二乘回归指数追踪模型通过了显著性检验同时解释了成分股和上证 50 指数之间绝大部分的线性关系。从表 7 的模型评估结果以及图 4 指数追踪图可以发现，各项模型评估指标值在三个模型中都是最小的，说明基于二进制粒子群特征选择算法结合最小二乘估计所得到的指数追踪模型的指数追踪效果最佳。

5.5. 成分股影响分析

为了到达利用少数的成分股便可实现对指数的追踪，即在使用更少的成本去投资少数几个优质股票尽可能获得接近于直接投资上证 50 指数所获得收益并降低投资风险的目的，需要讨论各成分股对于上证 50 指数的影响程度。最终是要提取出能够准确及时反映上证 50 指数日收盘价趋势的成分股即所谓的优质股票。

成分股对指数日收盘价波动大小的影响，通常会使用回归系数来衡量，但回归系数仅仅只描述了成分股日收盘价对指数日收盘价贡献的权重大小。一只高回归系数的成分股可以表明在指数日收盘上有更大的权重作用，但如果该成分股的波动变化比较小，那么最后对于指数日收盘价波动的组成占比也不会太大。所以在考虑回归系数的基础上同样还考虑成分股方差的大小，两者结合来衡量成分股对指数收盘价波动的影响。本文将使用以下公式计算成分股对指数日收盘价波动的影响大小：影响系数 = 回归系数*成分股日收盘价方差，为了更好的比较不同模型下各个成分股对指数日收盘价的影响，再对影响系数进行 0~1 区间标准化，同时为了更好的展示各成分股的影响系数，对各成分股的影响系数同时乘上 100。

基于三个指数追踪模型计算各成分股影响系数如下：

基于岭回归模型计算各成分股影响系数如表 8 所示：

Table 8. The influence coefficient of component stocks in ridge regression model
表 8. 岭回归模型的成分股影响系数

股票名称	影响系数	排名	股票名称	影响系数	排名
贵州茅台	100.00	1	伊利股份	2.53	14
山西汾酒	21.14	2	中信证券	2.49	15
中国中免	19.25	3	上汽集团	2.34	16
药明康德	14.62	4	中信建投	2.26	17
兆易创新	14.26	5	北方稀土	2.21	18
闻泰科技	11.80	6	兴业银行	2.18	19
华友钴业	8.70	7	通威股份	2.14	20
恒瑞医药	6.49	8	海尔智家	1.57	21
万华化学	5.90	9	工商银行	1.30	22
中国平安	5.20	10	中国石化	1.22	23
长城汽车	4.67	11	保利发展	1.17	24
海螺水泥	2.87	12	天合光能	0.40	25
航发动力	2.78	13	中国神华	0.00	26

虽然在构建岭回归模型的过程中已经剔除了不显著的变量，但是从上表还是可以发现某些成分股对于上证 50 指数日收盘价的走势不够敏感，比如中国神华、天合光能、保利发展和中国石化这些股票的影响系数是远远小于排名靠前的股票影响系数。

基于 Lasso 回归模型计算各成分股影响系数如表 9 所示：

Table 9. The influence coefficient of component stocks in lasso regression model
表 9. Lasso 回归模型的成分股影响系数

股票名称	影响系数	排名	股票名称	影响系数	排名
贵州茅台	100.00	1	中信证券	1.49	14
药明康德	14.38	2	闻泰科技	0.77	15
海天味业	10.89	3	兴业银行	0.74	16
兆易创新	8.46	4	上汽集团	0.66	17
山西汾酒	6.30	5	三一重工	0.65	18
中国平安	5.27	6	华泰证券	0.35	19
长城汽车	3.90	7	保利发展	0.18	20
恒瑞医药	3.76	8	航发动力	0.09	21
华友钴业	2.77	9	国电南瑞	0.04	22
天合光能	2.43	10	中国石油	0.01	23
通威股份	2.19	11	中国建筑	0.01	24
万华化学	1.75	12	包钢股份	0.00	25
海螺水泥	1.62	13			

虽然在构建 Lasso 回归模型的过程中已经剔除了不显著的变量，但是从上表还是可以发现某些成分股对于上证 50 指数日收盘价的走势不够敏感，比如包钢股份、中国建筑、中国石油和国电南瑞这些股票的影响系数是远远小于排名靠前的股票影响系数。

基于最小二乘回归模型计算各成分股影响系数如表 10 所示：

Table 10. The influence coefficient of component stocks in least squares regression model
表 10. 最小二乘回归模型的成分股影响系数

股票名称	影响系数	排名	股票名称	影响系数	排名
贵州茅台	100.00	1	恒力石化	0.97	18
韦尔股份	50.71	2	三一重工	0.96	19
中国中免	12.88	3	中信建投	0.85	20
片仔癀	7.19	4	华泰证券	0.64	21
药明康德	6.85	5	兴业银行	0.62	22
海天味业	4.84	6	伊利股份	0.61	23
招商银行	4.47	7	上汽集团	0.61	24
恒瑞医药	4.32	8	中国太保	0.56	25
中国平安	3.72	9	紫金矿业	0.38	26
隆基绿能	3.55	10	中远海控	0.36	27
长城汽车	2.31	11	保利发展	0.36	28
天合光能	2.05	12	长江电力	0.35	29
复星医药	2.00	13	中国建筑	0.27	30
通威股份	1.69	14	三峡能源	0.26	31
中国神华	1.35	15	中国石化	0.24	32
航发动力	1.34	16	国电南瑞	0.00	33
万华化学	1.13	17			

虽然在构建最小二乘回归模型的过程中已经剔除了不显著的变量，但是从上表还是可以发现某些成分股对于上证 50 指数日收盘价的走势不够敏感，比如国电南瑞、中国石化、三峡能源和中国建筑这些股票的影响系数是远远小于排名靠前的股票影响系数。

从上述三个模型所得到的成分股影响系数结果来看，各个模型所得到的结果都有所差异，但有一个共同点就是贵州茅台的影响系数在三个模型结果中都是最高的。

5.6. 投资行为模拟及收益分析

在上述内容中利用各个回归模型提取出了优质成分股影响排名的情况下，本文将在此基础上进行投资行为模拟和收益分析，以判断上述所得到的优质成分股是否能够达到本文想要研究的目的。

在投资行为中，有一个比较简单的核心思想：在暴跌之前卖出，在暴涨之前买入。基于以上思想本文主要以图 5 中所示的上证 50 指数波动比较大的三个区间来进行投资行为模拟。

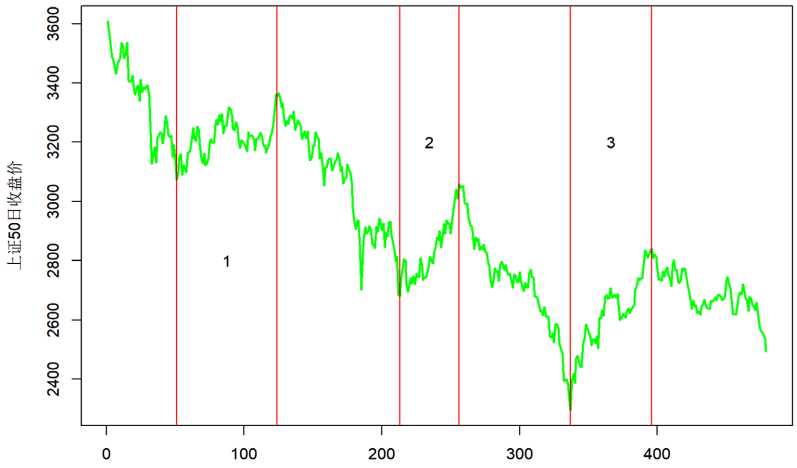

Figure 5. Investment interval
图 5. 投资区间

图 5 标示出了[51, 124]、[213, 256]、[337, 396]三个区间，[51, 124]区间表示样本 51 (时间 2021/8/23)到样本 124 (时间 2021/12/10)这段范围，这三个区间都是上证 50 指数暴跌后强势反弹的一个区间。本文将以这三个区间为例，分析投资优质成分股票和直接投资上证 50 指数所带来的收益对比。

现给出三个前提条件：

1) 以 50 万为投资金额。

2) 可以以分数的形式购买相应份额的股票(各成分股投资金额是基于影响系数进行分配，所以不满足购买整数份额的股票)

3) 成分股投资上，选择影响排名前 K (K = 5, 10, 15, 20)四种情况的成分股进行投资

在以上三个前提条件下给出以下模拟步骤：

1) 资金分配：以各成分股的影响系数为比重分配 50 万资金为各自成分股投资资金；以 50 万资金作为上证 50 的投资资金

2) 买入：将各成分股和上证 50 的投资资金除以上述 3 个区间的初始时刻(上证 50 日收盘价处于谷底)各自日收盘价得到各个成分股和上证 50 的买入份额

3) 卖出：将第三步所获得的各成分股和上证 50 的份额乘以上述 3 个区间的末时刻(上证 50 日收盘价处于顶点)各自日收盘价得到各个成分股和上证 50 的收益。

4) 将各个成分股所得收益求和得到投资成分股所得收益并减去原始投资资金 50 万得到最终净收益，与直接投资上证 50 指数所得净收益进行比较。

经过上述的模拟之后得到三个模型下的投资结果如表 11 所示：

Table 11. Income statement
表 11. 收益表

投资影响排名前 5 成分股				
投资区间	岭回归	Lasso 回归	最小二乘回归	上证 50
[51, 124]	127496.5	141956.4	134331.8	45705.0
[213, 256]	96456.8	77136.2	106893.1	70377.8
[337, 396]	186918.6	189689.2	183353.8	118296.8
总和	410871.9	408781.7	424578.7	234379.6

Continued

投资影响排名前 10 成分股				
投资区间	岭回归	Lasso 回归	最小二乘回归	上证 50
[51, 124]	116398.5	132413.1	123531.3	45705.0
[213, 256]	107626.6	88726.1	102522.1	70377.8
[337, 396]	175222.4	179110.8	179649.4	118296.8
总和	399247.5	400250.0	405702.8	234379.6
投资影响排名前 15 成分股				
投资区间	岭回归	Lasso 回归	最小二乘回归	上证 50
[51, 124]	111743.8	126832.6	122099	45705.0
[213, 256]	110249.8	91578.5	106547.4	70377.8
[337, 396]	170196.1	174253.6	173091	118296.8
总和	392189.7	392664.7	401737.3	234379.6
投资影响排名前 20 成分股				
投资区间	岭回归	Lasso 回归	最小二乘回归	上证 50
[51, 124]	110338.2	125236.9	119023.5	45705.0
[213, 256]	110980.7	91160.1	107373.4	70377.8
[337, 396]	167508.8	172884.6	170886	118296.8
总和	388827.7	389281.6	397282.9	234379.6

注：红色代表横向数据最大值(净收益收益最大)。

从表 11 可以得出如下结论：

1) 从投资净收益的角度来看，三个指数追踪模型分别所提取的影响排名前 5、10、15 和 20 优质成分股在三个区间的投资净收益都高于直接投资上证 50 指数所获得的净收益。

2) 基于岭回归、Lasso 回归和最小二乘回归所提取的优质成分股中，投资影响排名越靠前的成分股所获得的净收益越多。

3) 无论是选择影响排名前 5、10、15 还是 20 的成分股进行投资时，基于最小二乘回归指数追踪模型所得到的净收益是三个指数追踪模型中最高的。

4) 选择影响排名前 5,、10、15 和 20 的成分股进行投资时，基于三个指数追踪模型所得到的净收益以及净收益之间的差额都呈现一定的递减趋势。

5) 就本文所模拟的投资行为而言，基于最小二乘回归指数追踪模型所得到的净收益平均高出直接投资上证 50 指数所得净收益约 180,000 元。

6. 总结与建议

6.1. 总结

本文研究分析了 2021 年 6 月 10 日至 2023 年 5 月 31 日上证 50 指数及其成分股日收盘价的情况，构建了各成分股与上证 50 指数之间的回归模型，利用回归模型探讨了各成分股对上证 50 指数趋势及波动的影响，提取了各成分股中的优质股票并进行影响排名，经过投资行为模拟和收益分析证实三个指数追踪模型各自所提取的优质成分股是有效的，能够在相同投入资金下获得比直接投资上证 50 指数更高的净

收益，并且在投资成分股的选择下具有更大的灵活性和风险性控制。

首先基于二进制粒子群优化算法结合岭估计、Lasso 估计和最小二乘估计进行成分股选择，筛选出了优质成分股，成功构建了关于上证 50 指数的岭回归指数追踪模型、Lasso 回归指数追踪模型和最小二乘回归指数追踪模型。三个指数追踪模型中最小二乘回归指数追踪模型的拟合能力、预测能力和稳定性都是最优秀的，岭回归指数追踪模型的拟合能力、预测能力和稳定性都是最差的，但三个指数追踪模型都表现出了非常优秀的指数追踪能力。

然后基于三个回归模型所提取的成分股利用影响系数衡量了各个成分股对于上证 50 指数波动的影响并进行了影响排名，三组回归模型所提取的成分股都有所差别，但是从影响排名来看，三组回归模型的成分股影响排名中，排名最高的都是贵州茅台。

最后利用三组成分股影响排名进行了投资行为模拟并分析了收益情况，三个指数追踪模型各自所提取的优质成分股是有效的，能够在相同投入资金下获得比直接投资指数更高的净收益。相同投资资金下，选择越多的优质成分股进行投资，其获得的净收益一般来说会越低，但是其风险也会越低。其中无论是从追踪效果的角度还是投资的角度出发，基于二进制粒子群特征选择算法结合最小二乘估计所得到的指数追踪模型都是最佳模型。

6.2. 建议

1) 可以选择最小二乘回归指数追踪模型所提取的优质股票进行投资，就模拟投资而言，其带来的净收益最高。同样也可以基于三个指数追踪模型进行组合投资。

2) 追求高收益情况下，可以选择排名靠前的少数几只优质成分股进行投资，相对的也会带来更高风险。

3) 追求稳定的话，可以选择更多的优质成分股进行投资以降低风险性。

4) 本文主要研究是提取上证 50 指数优质成分股及其影响排名，在可靠的消息和经验支持下在未来一段时间上证 50 指数有一定的上涨趋势，再结合本研究所提取的优质成分股可能才会取得预期的收益。

5) 本研究旨在分析各成分股与上证 50 指数之间的关系以及提取优质成分股，为关于上证 50 指数的投资提供科学合理和可靠的方向和策略。并非单靠这份简单的分析报告就可以获得预期的收益，应当学习相关知识以及在大量经验和可靠信息下才能涉及投资领域。

参考文献

[1] Strub, O. and Baumann, P. (2018) Optimal Construction and Rebalancing of Index-Tracking Portfolios. *European Journal of Operational Research*, **264**, 370-387. https://doi.org/10.1016/j.ejor.2017.06.055

[2] 高见, 杨丹. 指数化投资中复制方法的比较分析[J]. 金融研究, 2006(8): 31-40.

[3] 陈晶. 股票指数非完全复制法及其再平衡策略的研究[D]: [硕士学位论文]. 上海: 复旦大学, 2012.

[4] 杨维, 李歧强. 粒子群优化算法综述[J]. 中国工程科学, 2004(5): 87-94.

[5] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[6] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[7] Kennedy, J. and Eberhart, R. (1995) Particle Swarm Optimization. *Proceedings of ICNN'95-International Conference on Neural Networks*, Perth, 27 November-1 December 1995, 1942-1948. https://doi.org/10.1109/ICNN.1995.488968

[8] Kennedy, J. and Eberhart, R.C. (1997) A Discrete Binary Version of the Particle Swarm Algorithm. *IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation*, Orlando, 12-15 October 1997, 4104-4108.

[9] 杨虎, 杨玥含. 金融大数据统计方法与实证[M]. 北京: 科学出版社, 2016: 122-123.

[10] 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. https://doi.org/10.27670/d.cnki.gcqdu.2019.000010