

基于邻近梯度算法求解Lasso回归模型在 证券指数的应用研究

王 琦, 彭定涛

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2024年9月13日; 录用日期: 2024年10月14日; 发布日期: 2024年11月13日

摘 要

在证券市场中, 指数跟踪是投资者进行资产配置和风险管理的重要手段。传统的指数跟踪方法往往有模型解释性差、计算复杂度高问题。因此, 探索新的算法和技术以提升指数跟踪的效果具有重要意义。本文考虑Lasso回归模型来进行变量选择, 首先介绍了Lasso回归模型基本原理, 然后利用邻近梯度算法求解回归系数, 该解具有稀疏性, 旨在众多的变量中精确选择有效的部分变量来预测证券指数。最后利用沪深300指数以及其成分股的收盘价格作为分析数据, 得出小部分股票就可以达到几乎相同的拟合效果, 通过实例说明该方法在证券指数跟踪中有一定的有效性和优越性, 能够实现更稀疏的变量组合, 给证券指数跟踪提供了新的思路和方法。

关键词

Lasso回归, 股票价格, 邻近梯度算法, 稀疏优化, 证券指数

Research on the Application of Lasso Regression Model Based on Proximity Gradient Algorithm in Stock Index

Qi Wang, Dingtao Peng

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 13th, 2024; accepted: Oct. 14th, 2024; published: Nov. 13th, 2024

Abstract

In the security market, index tracking is an important means for investors to allocate assets and

文章引用: 王琦, 彭定涛. 基于邻近梯度算法求解 Lasso 回归模型在证券指数的应用研究[J]. 电子商务评论, 2024, 13(4): 2973-2981. DOI: 10.12677/ecl.2024.1341482

manage risks. Traditional exponential tracking methods often have the problems of poor model interpretation and high computational complexity. Therefore, it is of great significance to explore new algorithms and techniques to improve the effect of exponential tracking. In this paper, Lasso regression model is considered for variable selection. Firstly, the basic principle of Lasso regression model is introduced, and then the adjacent gradient algorithm is used to solve the regression coefficient. The solution is sparse, aiming at accurately selecting effective part of the variables to predict the stock index from many variables. Finally, using the closing prices of CSI 300 index and its component stocks as analysis data, a small number of stocks can achieve almost the same fitting effect. An example shows that the method has certain effectiveness and superiority in securities index tracking, and can realize sparser variable combination, which provides a new idea and method for securities index tracking.

Keywords

Lasso Regression, Stock Price, Proximity Gradient Algorithm, Sparse Optimization, Stock Index

Copyright © 2024 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

在当今复杂多变的金融市场中, 股票分析一直是投资者、分析师以及学者关注的焦点。随着大数据和机器学习技术的飞速发展, 传统的股票分析方法正逐渐被更高效、更智能的模型所取代。其中, Lasso 回归作为一种强大的统计学习工具, 正逐渐成为股票分析领域的一颗新星。

Lasso 回归, 全称为 Least Absolute Shrinkage and Selection Operator (最小绝对收缩和选择算子), 与传统的线性回归相比, 模型中加入了惩罚项, 是一种能够在拟合模型的同时自动进行特征选择的方法, 通过调整 Lasso 回归中的正则化参数, 将那些对目标变量预测贡献不大的特征系数压缩到零, 从而有效减少模型的复杂度, 提高模型的泛化能力和预测精度[1]。

在股票分析领域, 特征选择尤为重要。因为股票市场受多种因素影响, 这些因素相互交织, 共同作用于股票价格的变化。然而, 并非所有因素都对股票价格具有显著影响, 因此过多的变量不仅会增加模型的计算复杂性, 还可能引入影响模型的噪声, 导致模型过拟合。因此, 通过 Lasso 回归进行特征选择, 筛选出对股票价格预测最关键的特征, 对于提高股票分析的准确性和效率都具有重要意义。

随着数据规模的增大, 求解 Lasso 问题需要更为先进的算法。最新的算法有梯度下降法、坐标下降法、交替方向乘法[2]和邻近梯度算法(Proximity gradient algorithm)等。本文主要介绍了邻近梯度算法, 该算法可以看作是投影梯度算法的推广。1962 年, Moreau 在投影算子中采用凸映射代替恒等映射, 引入了邻近算子的概念, 将投影梯度算法推广为邻近梯度算法, 具体是将目标函数中可微的部分进行二阶展开, 并保留非光滑部分, 然后求解极小值作为下一步的估计。这种方法特别适用于求解非光滑约束优化问题。本文旨在利用邻近梯度算法求解 Lasso 回归的系数, 通过实证分析展示其在提高股票预测精度和效率方面的作用, 为投资者和分析师提供一种更加智能化的股票分析算法[3]。

2. 预备知识

2.1. Lasso 的基本思想

Breiman 在 1995 [4]年提出 Nonnegative Garrote 方法, 是用来解决变量选择的回归模型。主要核心思

想是对数据集 $(x_i, y_i), i=1, 2, \dots, n$ 作回归拟合, 其中 $x_i \in R^s$, $y_i \in R$, 即 $x_i = (x_{i1}, x_{i2}, \dots, x_{is})$ 为自变量, y_i 为因变量。变量选择即从所有的变量因子中选择一小部分作为回归变量, 而这部分回归变量构成的回归函数也可以让我们有效地预测结果。由于每增加一个回归因子, 回归方程的方差都会增加, 因此变量选择变得很有必要, 可以减小估计方差的同时又简化回归模型。

而斯坦福大学统计学教授 Robert Tibshirani 于 1996 [5] 年在 Nonnegative Garrote 方法的基础上首次提出 Lasso 回归方法, 对于上述的数据集 (x_i, y_i) , 假设是 x_{ik} 标准化的, 即:

$$\sum_i \frac{x_{ik}}{N} = 0, \sum_i \frac{x_{ik}^2}{N} = 1.$$

考虑如下线性模型:

$$y_i = \sum_{k=1}^s \alpha_k x_{ik} + \varepsilon_i, i=1, 2, \dots, n,$$

写成矩阵形式为:

$$Y = X\alpha + \varepsilon.$$

这里, 被解释变量 Y 是 $n \times 1$ 的向量, 解释变量 X 是 $n \times s$ 的矩阵, 而我们得到误差项 ε 是 $n \times 1$ 的向量。Lasso 估计为:

$$\min \sum_i \left(y_i - \sum_k \alpha_k x_{ik} \right)^2, \text{ s.t. } \sum \alpha_k \leq s.$$

其中 $\alpha_k \geq 0, k=1, 2, \dots, N$ 是估计参数。

等价表示为标准的 Lasso 问题:

$$\min_{\alpha} F(\alpha) = \frac{1}{2} \|A\alpha - b\|_2^2 + \lambda \|\alpha\|_1, \quad (1)$$

其中 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_s)$, $\alpha_i \in R$, $\lambda > 0$ 为正则化参数, $\|\alpha\|_1 = \sum_{i=1}^s \alpha_i$, A 记为解释变量构成的矩阵, b 为被解释变量构成的向量。可以看出 Lasso 模型还是以最小二乘法计算残差, 但引入了 L1 范数作为惩罚项用来减少模型的回归系数。

2.2. 邻近算子

定义 2.2.1: [6] 设 ψ 是 R^n 上的实值凸函数。对于任意给定的 $t > 0$, ψ 的邻近映射定义为:

$$\text{prox}_{t\psi}(x) := \arg \min_{z \in R^n} \left\{ \psi(z) + \frac{1}{2t} \|z - x\|_2^2 \right\}, \forall x \in R^n.$$

称映射 $\text{prox}_{\lambda\psi} : R^n \rightarrow R^n$ 为 $\lambda\psi$ 的邻近算子。

对于 $a \in R$, $\lambda > 0$, L1 范数的邻近算子为:

$$\begin{aligned} \text{prox}_{\lambda\|\cdot\|_1}(a) &= \text{sign}(a) * \max\{|a| - \lambda, 0\} \\ &= \begin{cases} a - \lambda, & a > \lambda \\ 0, & 0 \leq a < \lambda \\ a + \lambda, & a < -\lambda \\ 0, & -\lambda \leq a < 0 \end{cases} \end{aligned} \quad (2)$$

3. 邻近梯度算法

求解问题(1)困难在于 $\|x\|_1$ 的不可微性, 这使得传统的光滑优化算法不能得到直接应用, 由此邻近算

法被 Moreau 于 1962 年提出, Rockafellar [7] 随后对邻近算子做出进一步研究, 并应用于优化问题。具体如下:

优化问题模型为:

$$\min_{x \in R^n} F(x) := f(x) + \lambda \|x\|_1, \quad (3)$$

其中 $f(x)$ 为损失函数, 本文使用最小二乘作为损失函数, 即 $f(x) = \frac{1}{2} \|Ax - b\|_2^2$, 则优化模型为:

$$\min_{x \in R^n} F(x) := \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1. \quad (4)$$

邻近梯度算法的迭代格式为:

$$x_{k+1} \in \text{prox}_{\gamma_k \lambda} (x_k - \gamma_k \nabla f(x_k))$$

其中 λ_k 为迭代步长, 步长可以通过线搜索得到。

邻近梯度算法的迭代格式简单, 适用于求解大规模问题, 能够处理非光滑且目标函数为广义实值函数的问题。邻近梯度算法广泛应用于图像处理、压缩传感、机器学习、系统识别、协同预测、低维嵌入、数据挖掘和模式识别等领域[8] [9]。

3.1. 求模型参数过程

算法 1: 迭代算法

邻近梯度算法

初始化: 设置 $x^{-1} = x^0$, $\delta > 0$, $\rho \in (0, 1)$, $k = 0$ 。

迭代过程:

计算

$$x_{k+1} \in \text{prox}_{\gamma_k \lambda} (x_k - \gamma_k \nabla f(x_k)) \quad (5)$$

如果满足:

$$F(x_{k+1}) \leq F(x_k) - \frac{\delta}{2} \|x_{k+1} - x_k\|^2,$$

则 $k = k + 1$, 否则设置 $\gamma_k = \gamma_k \rho$ 并回(5)。

停机准则:

$$\frac{\|x_{k+1} - x_k\|_2}{\max\{1, \|x_{k+1}\|_2\}} < 10^{-4} \text{ 或者达到最大的迭代次数(Maxiter).}$$

3.2. 邻近梯度算法的收敛性

定义 3.2.1 (P 稳定点) [10]: 存在一个 $\gamma > 0$ 使得

$$x^* \in \text{prox}_{\gamma \lambda} (x^* - \gamma \nabla f(x^*)) \quad (6)$$

则称 x^* 为问题(3)的 P 稳定点。

引理 3.2.2: 设序列 $\{x_k\}$ 由邻近梯度算法产生, 则:

(i) $\sum_{n=1}^{+\infty} \|x_{n+1} - x_n\|^2 \leq +\infty$;

(ii) 函数值序列 $\{F(x_k)\}$ 具有下降性质, 即 $\{F(x_k)\}$ 是非递增的。

证明:

迭代停机准则知, 迭代最多到 100 次, k 有上限, 记 N 为 k 值的集合。从算法中可知:

$$\frac{\delta}{2} \|x_{k+1} - x_k\|^2 \leq F(x_k) - F(x_{k+1})$$

$k = 0, 1, \dots, K$, 累加有:

$$\sum_{k=0}^K \frac{\delta}{2} \|x_{k+1} - x_k\|^2 \leq F(x_0) - F(x_{K+1}),$$

令 K 趋于 $+\infty$, 可得(i)。

由算法的迭代过程可知, 产生的 x_{k+1} 要满足式子 $F(x_{k+1}) \leq F(x_k) - \frac{\delta}{2} \|x_{k+1} - x_k\|^2$, 因此有:

$$F(x_{k+1}) - F(x_k) \leq -\frac{\delta}{2} \|x_{k+1} - x_k\|^2 \leq 0,$$

$$F(x_{k+1}) \leq F(x_k),$$

说明 $\{F(x_k)\}$ 是非递增的。

定理 3.2.3: 邻近梯度算法产生的序列 $\{x_k\}$ 收敛到 P 稳定点。

证明: 设 x^* 为序列 $\{x_k\}$ 的累积点, 即当 k 趋于 $+\infty$, 有 $x_k \rightarrow x^*$, 由迭代算法可知:

$$x_{k+1} \in \text{prox}_{\gamma_k \lambda \|\cdot\|_1} (x_k - \gamma_k \nabla f(x_k)),$$

两边同时取极限有:

$$x^* \in \text{prox}_{\gamma \lambda \|\cdot\|_1} (x^* - \gamma \nabla f(x^*)),$$

根据定义 3.2.1 知 x^* 为 P 稳定点, 因此序列 $\{x_k\}$ 收敛到 P 稳定点。

4. 基于邻近梯度算法用 Lasso 回归对沪深 300 指数回归拟合

金融市场是经济体系的重要组成部分, 其发展与宏观经济状况密切相关。经济增长、通货膨胀、利率变动等经济因素直接影响金融市场的走势。而金融市场指数是衡量特定金融市场或行业整体表现的重要工具, 评判市场健康程度、活跃度以及经济发展状况的重要依据, 它综合了众多相关资产的价格变动, 为投资者和分析师提供了市场走势的直观参考。

主要的金融市场指数包括股票市场指数、债券市场指数和商品市场指数等。这些指数通过计算样本股票的价格平均值或市值加权平均值, 来反映市场的总体走势和估值水平。本文以沪深 300 指数为例进行数据拟合试验, 沪深 300 指数由证监会于 2005 年 4 月 8 日正式推出, 是中国 A 股市场的一个重要基准指数, 它涵盖了沪深两市总市值最大、流动性最好的 300 只股票, 涵盖了金融、消费、工业、能源等多个行业, 具有广泛的市场代表性, 这些行业之间的关联性和差异性为 Lasso 回归提供了丰富的变量选择空间。同时沪深 300 指数及其成分股的历史数据丰富, 这为使用 Lasso 回归进行建模提供了坚实的基础。

但沪深 300 指数的构成复杂且权重为动态权重, 因此, 构建一个稀疏的组合对指数进行拟合来提高效率。我们运用 Lasso 方法对沪深 300 指数作回归拟合, 从沪深 300 的成分股中找出小部分股票就可以达到几乎相同的拟合效果。在此过程中, 我们用邻近梯度算法进行优化迭代。

本文选取数据从 2024 年 7 月 8 日至 2024 年 9 月 6 日, 共计 45 个交易日的数据, 在各个价格数据中, 收盘价代表了一天交易结束的买卖方的均衡价格, 最具有代表性以及信息量, 所以选取收盘价作为拟合价格, 包含沪深 300 指数及其 300 支成分股的所有收盘价数据。数据来源为 Choice 金融终端。

4.1. 基于邻近梯度算法的 Lasso 问题实例

我们用邻近梯度算法进行优化迭代, 设第 j 支股票在第 i 天的收盘价为 x_{ij} , 记矩阵 $A = (x_{ij})$, $i = 1, \dots, n$, $j = 1, \dots, m$, 第 i 天的沪深 300 指数的收盘价为 b_i , $i = 1, \dots, n$ 。则 Lasso 模型为:

$$\min \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1.$$

4.1.1. 数据预处理

在 Lasso 方法中, 均假设数据是标准化的, 即 $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 0$ 。因此, 我们首先要对每一支股票的收盘价格和沪深 300 指数的收盘价进行标准化处理, 即令 $\hat{x}_{ij} = \left(x_{ij} - \sum_i x_{ij}/N\right) / \text{std}(x_{ij})$, $\hat{b}_i = \left(b_i - \sum_i b_i/N\right) / \text{std}(b_i)$, 其中 $\text{std}(x_{ij})$ 为 $(x_{1j}, x_{2j}, \dots, x_{Nj})^T$ 的标准差, 则标准化后的模型为:

$$\min \frac{1}{2} \|\hat{A}x - \hat{b}\|_2^2 + \lambda \|x\|_1.$$

通过求解此惩罚回归问题, 希望通过邻近梯度算法构建其成分股的稀疏回归, 数学表达式如下:

$$\hat{b} - \sum_T \hat{\beta}_k x_k, T \subseteq \Omega \text{ 且 } |T| \leq |\Omega|.$$

其中 Ω 是全部股票的指标集, T 为回归系数非零的股票指标集合, x_k 为回归系数, $\hat{\beta}_k$ 为各成分股的收盘价。

4.1.2. 数据分析

在收集到数据后用 MATLAB 进行处理, 所有数值实验均使用 MATLAB (R2018b) 在联想 PC 上进行 (Intel (R) Core (TM) i5-9500 CPU (3.00 GHz), 8.00 GB (RAM))。

参数设定为 $\lambda = 8$, $\gamma_0 = 0.01$, $\rho = 0.5$, $\text{maxiter} = 100$, (最大迭代次数), 在此参数下, 对数据进行拟合并分析。从图 1 可以看出, 大多数成分股的回归系数为 0, 只有 28 个非零的回归系数, 从而达到了稀疏的目的, 可以从这 28 个成分股来拟合沪深 300 指数。

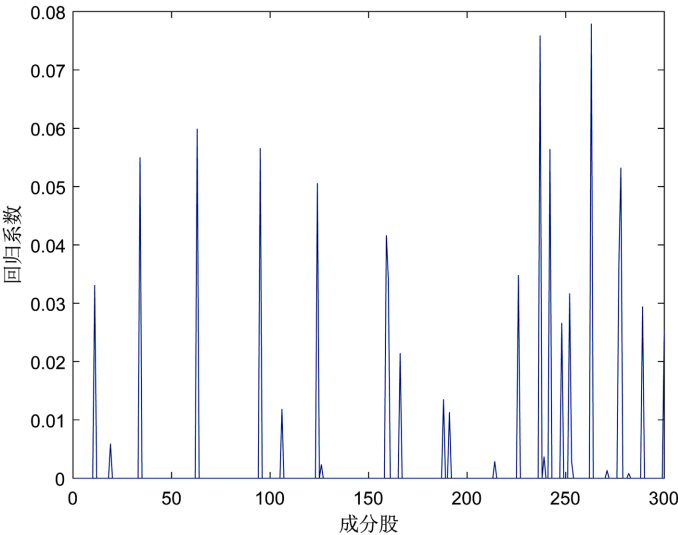


Figure 1. Component regression coefficient
图 1. 成分股回归系数

列出了回归系数不为零的 28 个成分股的股票代码，名称和回归系数，见表 1。

Table 1. Names of constituent stocks and regression coefficients at $\lambda = 8$

表 1. $\lambda = 8$ 时的成分股名称以及回归系数

股票代码	股票名称	回归系数	股票代码	股票名称	回归系数
300418	昆仑万维	0.0330	000625	长安汽车	0.0058
600489	中金黄金	0.0549	600016	民生银行	0.0598
688008	澜起科技	0.0565	600085	同仁堂	0.0118
002142	宁波银行	0.0505	603195	公牛集团	0.0023
600887	伊利股份	0.0415	000157	中联重科	0.0340
002415	海康威视	0.0213	000063	中兴通讯	0.0134
002371	北方华创	0.0112	000977	浪潮信息	0.0028
002241	歌尔股份	0.0347	600346	恒力石化	0.0758
002648	卫星化学	0.0036	603501	韦尔股份	0.0563
603986	兆易创新	0.0265	600584	长电科技	0.0316
600745	闻泰科技	0.0028	000568	泸州老窖	0.0778
000786	北新建材	0.0013	300782	卓胜微	0.0370
000938	紫光股份	0.0531	600183	生益科技	0.0008
002916	深南电路	0.0293	600426	华鲁恒升	0.0293

此参数下稀疏拟合回归和原始数据的对比如图 2，可以看出两条曲线差值不大，最大的差值为 0.5639，说明拟合效果理想，两条曲线是趋于重合的。

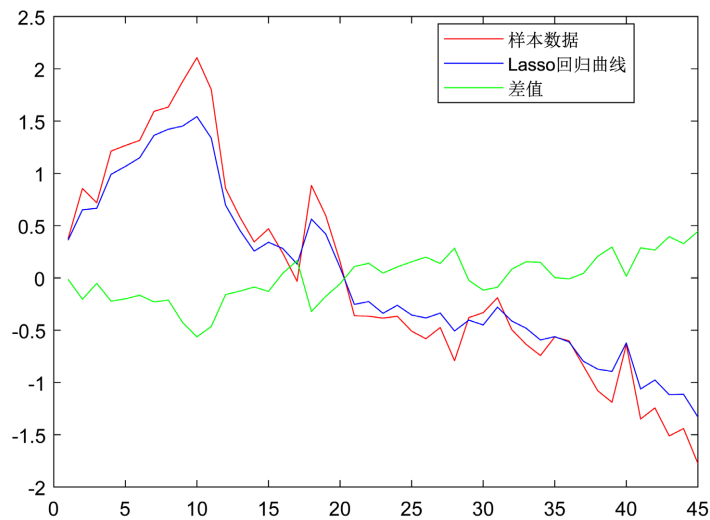


Figure 2. Regression fitting versus raw data

图 2. 回归拟合和原始数据对比

同时还采用控制变量法，分析了不同的参数对于拟合结果的影响。对于不同的参数 λ 下得到的回归

系数非零个数不同, 如图 3, 可以看出惩罚因子 λ 增大非零系数个数会减少, λ 在 0 到 5 之间有一个相对大的减小。

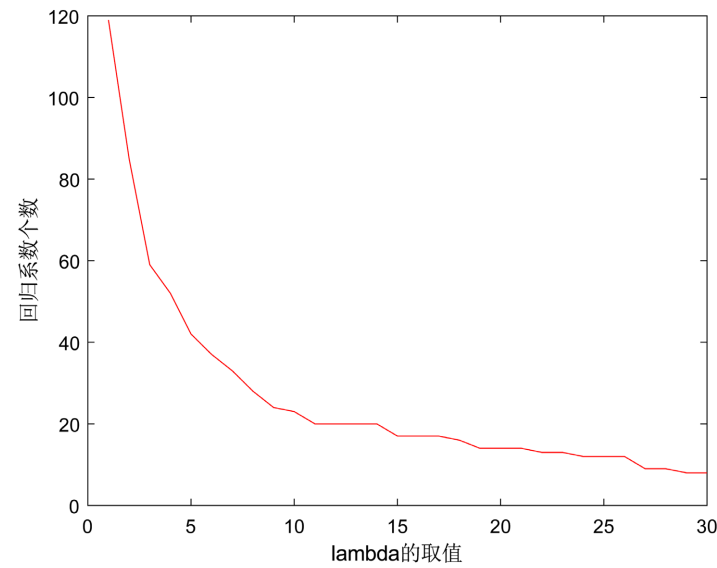


Figure 3. Different corresponding number of non-zero regression coefficients

图 3. 不同 λ 对应的非零回归系数个数

在其他参数不变的情况下, 设置 $\lambda=15$ 回归系数的非零个数为 17 个, 成分股名称在表 2:

Table 2. Stock code and name at $\lambda=15$

表 2. $\lambda=15$ 时的成分股代码和名称

股票代码	股票名称	股票代码	股票名称
300418	昆仑万维	600489	中金黄金
600016	民生银行	688008	澜起科技
002142	宁波银行	600887	伊利股份
002415	海康威视	000063	中兴通讯
002241	歌尔股份	600346	恒力石化
603501	韦尔股份	603986	兆易创新
000568	泸州老窖	300782	卓胜微
000938	紫光股份	600183	生益科技
002916	深南电路		

从表中可以看出这 17 个也在 $\lambda=8$ 时的 28 个非零回归系数对应的成分股中, 说明随着的 λ 增大, 剩下的变量对拟合结果重要性更高, 不容易被惩罚掉。而保留影响相对较大的变量, 它们对于结果的相关性更高, 对拟合效果的贡献更大。

5. 总结

在证券市场中, 构建一种高效的指数跟踪资产组合, 以紧密跟随目标指数的表现, 对于无论是个人还是机构投资者而言, 都具有重要的实践意义。本文致力于探索并应用一种创新的方法, 即通过模型和变量选择技术来寻找稀疏解, 特别是采用邻近梯度算法来求解 Lasso 回归模型, 本文实现了仅利用少数精选股票来拟合并复现一个原本包含数百支股票的复杂指数指标, 从而构建出一个稀疏而高效的指数跟踪资产组合。这种方法为投资者提供了一种更为灵活且成本效益高的指数复制和跟踪策略, 通过这种方式, 投资者不仅能够显著降低投资成本, 还能够有效分散风险, 并最终提升整体的投资效率。这一研究成果为证券市场中的指数跟踪实践提供了新的视角和有力的工具支持。

基金项目

国家自然科学基金(12261020), 贵州省科技计划(黔科合基础 ZK [2021]一般 009)和贵州省高层次留学人才创新创业择优资助重点项目([2018] 03)。

参考文献

- [1] Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, **7**, 2541-2563.
- [2] 何宜盛. Lasso 问题以及其在证券指数稀疏回归中的应用[D]: [硕士学位论文]. 南京: 南京大学, 2016.
- [3] 郑家萌. Lasso 回归模型在证券指数中的应用研究[D]: [硕士学位论文]. 广州: 暨南大学, 2018.
- [4] Breiman, L. (1995) Better Subset Selection Using the Non-Negative Garotte. *Technometrics*, **37**, 373-384. <https://doi.org/10.1080/00401706.1995.10484371>
- [5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- [6] Shen, H.L., Peng, D.T. and Zhang, X. (2024) Smoothing Composite Proximal Gradient Algorithm for Sparse Group Lasso Problems with Nonsmooth Loss Functions. *Journal of Applied Mathematics and Computing*, **70**, 1887-1913. <https://doi.org/10.1007/s12190-024-02034-2>
- [7] Rockafellar, R.T. (1976) Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, **14**, 877-898. <https://doi.org/10.1137/0314056>
- [8] Bian, W. and Chen, X.J. (2020) A Smoothing Proximal Gradient Algorithm for Non-Smooth Convex Regression with Cardinality Penalty. *SIAM Journal on Numerical Analysis*, **58**, 858-883. <https://doi.org/10.1137/18M1186009>
- [9] 刘倩. 极小化两个函数和的邻近梯度法研究[D]: [硕士学位论文]. 成都: 西华大学, 2017.
- [10] Zhou, S.L. (2022) Gradient Projection Newton Pursuit for Sparsity Constrained Optimization. *Applied and Computational Harmonic Analysis*, **61**, 75-100. <https://doi.org/10.1016/j.acha.2022.06.002>