Operations Research and Fuzziology 运筹与模糊学, 2023, 13(4), 3527-3535
Published Online August 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.134355

Hans汉斯

基于分位数回归的两步估计稀疏指数追踪

马　林

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年6月10日；录用日期：2023年8月5日；发布日期：2023年8月11日

摘　要

指数追踪这种被动管理方法凭借其风险小成本低的优势，受到大量投资者的追捧，其目标是最小化追踪误差。本文采用部分复制的策略对上证180指数进行指数追踪，以均方根误差作为衡量标准。利用表现较好的两步估计方法来追踪指数，第一步利用弹性网进行变量选择，第二步考虑模型的稳健性使用分位数回归来确定系数。实证分析结果显示两步估计方法优于单一方法，而基于分位数回归两步估计是其中表现最好的，因此可以使用此模型来进行指数追踪。

关键词

指数追踪，稳健估计，分位数回归，两步估计

Two-Step Estimation Sparse Index Tracking Based on Quantile Regression

Lin Ma

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jun. 10th, 2023; accepted: Aug. 5th, 2023; published: Aug. 11th, 2023

Abstract

Index tracking is a passive management method that is sought after by a large number of investors due to its low risk and low cost, and its goal is to minimize tracking errors. In this paper, the index tracking of the SSE 180 Index is carried out by a partially replicated strategy, and the root mean square error is used as the measurement standard. The index is tracked using a well-performing two-step estimation method, the first step using the elastic net for variable selection, and the second step considering the robustness of the model using quantile regression to determine the coefficients. Empirical analysis results show that the two-step estimation method is superior to the single-step method, and the two-step estimation based on quantile regression is the best, so this model can be used for exponential tracking.

文章引用: 马林. 基于分位数回归的两步估计稀疏指数追踪[J]. 运筹与模糊学, 2023, 13(4): 3527-3535.
DOI: 10.12677/orf.2023.134355

Keywords

Index Tracking, Robust Estimation, Quantile Regression, Two-Step Estimation

1. 引言

1.1. 研究背景

证券市场作为一个重要的投资途径，一直以来受到大量投资者的追捧和喜爱，同时还兴起了各种投资方法。金融投资管理策略可以分为主动投资管理和被动投资管理，主动投资为了获得一个超出特定基准(如股票指数)的回报，要求投资者能够准确预测证券市场的走向，看准时机买入或卖出以获得高额的利润，主动投资需要投资者频繁的进行交易，因此会产生额外的交易和管理费用并且主动投资风险较大；被动投资则是以长期收益和有限管理为出发点来进行投资，相较于主动投资风险较小成本较低，而指数追踪(index tracking)正是一种流行的被动投资。

指数追踪也称为指数复制(index replication)，需要从指数中选定特定的股票资产并按照一定策略进行资产分配，复制指数的表现，从而得到和目标指数相近的收益。指数追踪可以分为完全复制(full replication)和部分复制(partial replication)，完全复制要求投资者使用指数池的所有股票来复制指数的表现，追踪效果很好，但当指数的成分股数量较多时，需要很高的成本和精力，可行性不高。而部分复制只使用指数池里的少部分股票就能得到一个比较满意的追踪效果，更加受到投资者的关注和青睐，因此本文采用部分复制的策略来进行稀疏指数追踪。

对于部分复制的指数追踪分为两个部分：1) 股票选择，即从指数池中选出优质的股票，后面部分都是基于所选股票来做的；2) 权重确定，即在选出来的这些股票中进行资产分配，每只股票按多少比例进行投资。股票选择或者权重确定部分处理不当都会对投资者收益影响很大，两个部分都很关键。

1.2. 研究现状

近年来随着机器学习的发展和兴起，国外许多学者开始使用传统机器学习和深度学习方法来对指数追踪进行研究。2020 年，Kim 等[1]使用深度自动编码器和堆叠自动编码器来选择股票，最后使用等权重的策略来确定权重。2021 年，Kwak 等[2]希望得到一个不随输入变化而变化的权重，提出一种固定噪声作为输入的神经网络来进行指数追踪，与大多数只使用机器学习来选择股票方法相比，他们提出的方法可以进行股票选择和权重确定。2022 年，Bradrania 等[3]考虑不同的市场状态使用不同选股标准，提出一种基于市场状态的股票选择方法。2022 年，Cao 等[4]使用随机森林来进行股票选择，并且考虑了多重共线性利用岭估计来确定权重。

2014 年，Lan 等[5]使用非负 lasso 来追踪沪深 300 (CSI300)指数，还提出仅使用非负 lasso 来选择股票和使用非负最小二乘来估计权重的两步估计，在实证研究中发现两步估计的追踪效果更好。2014 年，Lan 等[6]使用非负弹性网来追踪沪深 300 指数和上证 180 (SSE180)指数，并证明了它的变量选择一致性，还提出仅使用非负弹性网来选择股票和使用非负最小二乘来估计权重的两步估计，发现两步估计的表现比直接使用非负弹性网好。2016 年 Yuehan 等[7]使用非负自适应 lasso 和非负自适应 lasso 加非负最小二

乘的两步估计来追踪沪深 300 指数，其中的自适应权重用于惩罚 L_1 惩罚项中不同的回归系数。

2021 年，Ning 等[8]结合自适应 L_1 惩罚和 L_2 惩罚提出非负自适应弹性网来追踪沪深 300 指数，证明了非负自适应弹性网的变量选择一致性。2021 年，Ning 等[9]使用 MCP 惩罚下的非负估计来进行指数追踪。2022 年，Qian 等[10]考虑时间对金融数据的影响提出了一种新的时间加权非负 lasso 模型，证明了变量选择一致性和渐进无偏性，进行指数追踪。

上述的非负 lasso、非负弹性网、非负自适应 lasso 和非负自适应弹性网都是基于最小二乘估计，对异常值很敏感，不是一种稳健的估计。为了得到一个稳健的指数追踪模型，许多学者做了大量工作。2020 年，Ning [11]结合复合分位数回归和 L_1 惩罚项来进行指数追踪，具有稳健性。2022 年，Ning [12]使用基于 HUB 和 LAD 的 M 估计来追踪标准普尔 500 指数。

本文的特色之处在于：1) 在指数追踪的研究中，发现使用两步估计的追踪表现往往会比使用单一方法的表现要好，因此本文使用一种新的两步估计来进行指数追踪。2) 大多关于指数的追踪研究都是基于最小二乘法，容易受到异常值的影响，不是一种稳健的方法，而本文提出的模型具有稳健性。

2. 两步估计法

为了保证构建的股票组合具有更好的指数追踪效果，本文使用两步估计法来进行指数追踪。

弹性网(Elastic Net)是一种用于线性回归的正则化方法，它结合了 L_1 正则化(Lasso)和 L_2 正则化(Ridge)的优点，可以同时选择变量和保留相关变量的特点。因此本文使用弹性网来选取股票，再考虑到稳健性使用分位数回归来估计股票权重。

对于经典的线性模型：

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

其中 $\boldsymbol{Y}$ 是一个 n 维向量， $\boldsymbol{X}$ 为 $n\times p$ 维矩阵， $\boldsymbol{\beta}=\left(\beta_1,\beta_2,\cdots,\beta_p\right)^{\mathrm{T}}$ 是一个 p 维向量， $\boldsymbol{\varepsilon}=\left(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\right)^{\mathrm{T}}$ 是一个 n 维独立同分布且独立于 $\boldsymbol{X}$ 的随机误差向量。

2.1. 弹性网

为了得到一个稀疏的模型来进行股票选择，本文使用基于 L_1 正则和 L_2 正则的弹性网，对于给定数据(X, Y)，弹性网的定义如下：

$$\hat{\beta}=\arg\min\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right\|_2^2+\lambda_1\left\|\boldsymbol{\beta}\right\|_1+\lambda_2\left\|\boldsymbol{\beta}\right\|_2^2$$

其中 λ_1 和 λ_2 是调优参数，当 $\lambda_2=0$ 时，弹性网就变成了 lasso。

弹性约束估计在处理高维数据时特别有用，可以帮助减少过拟合和选择最相关的特征。在弹性约束估计中，目标是最小化损失函数，该函数由平方误差项和带有 L_1 和 L_2 惩罚项的正则化项组成。L_1 正则化项通过将系数的绝对值相加来惩罚模型中的不相关变量，L_2 正则化项通过将系数的平方和相加来惩罚模型中的系数过大。通过调整 L_1 和 L_2 正则化项之间的权重，可以控制变量的选择和模型的复杂度。

2.2. 分位数回归

分位数回归(Quantile Regression, QR)是一种统计方法，用于估计输入变量和输出变量之间的条件分位数，是一种稳健估计。与普通最小二乘法(Ordinary Least Squares, OLS)回归不同，它不仅关注因变量的均值与自变量之间的关系，而且关注因变量的整个分布。分位数回归能提供更多关于数据的信息，尤其在因变量的分布具有偏斜性或存在异常值的情况下。

分位数回归的目标是最小化以下的损失函数：

$$\text{argmin}_{\beta}\sum_{i=1}^{n} p_q\left(y_i - x_i\beta(q)\right)$$

其中，$\beta(q)$ 表示在 q 分位数下自变量对因变量的影响系数，p_q 为分位数损失函数，可以定义为：

$$p_q(u) = u(q)\cdot\left(q - I(u<0)\right)$$

其中 $u(q)$ 表示 u 在 q 分位数下的值，$I(u<0)$ 是示性函数，当 $u<0$ 时取值为 1，否则为 0。

3. 实证分析

本文的数据来源于 choice 金融终端，选用上证 180 指数收盘价日线数据，按照训练集: 测试集 = 4:1 将样本进行划分。运用两步估计的方法，第一步利用弹性网进行变量筛选，第二步选用分位数回归模型来确定系数，得到最终的指数追踪模型。为了体现本文所提方法的追踪表现，与弹性网以及其他两步估计法(如弹性网 + ols、弹性网 + 主成分回归)作对比。

由于收盘价的数值较大，本文选用均方根误差(RMSE)作为模型评估的衡量标准，其定义如下：

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} = \sqrt{\text{MSE}}$$

3.1. 弹性网

首先利用弹性网来进行变量选择，先进行 CV 交叉验证来确定调优参数 λ_1 和 λ_2，进一步得到保留的变量个数为 32。

直接使用弹性网估计所选 32 个变量的系数，然后计算出拟合值与预测值并绘制指数追踪图，如图 1 所示，黑线、红线、蓝线分别表示指数的实际值、拟合值、预测值，可以看出弹性网在训练集的表现很好，算得训练集上的 RMSE 为 18.872；但在测试集上弹性网不能很好的同步指数实际值的走势，RMSE 为 188.201，比在测试集的效果相差很多。

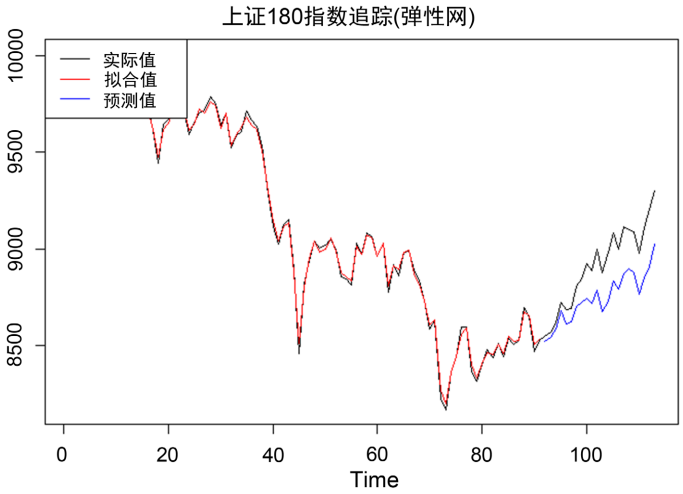

Figure 1. Diagram of the index tracking of elastic net
图 1. 弹性网指数追踪图

3.2. 弹性网 + 普通最小二乘

接下来准备两步估计来进行指数追踪，即仅用弹性网进行变量选择，再用普通最小二乘回归来确定系数，在这之前先对残差进行相关的检验。

如图 2 所示，散点分布虚线所在的直线周围，并且对残差进行 W 检验，求得 P 值为 0.7617，残差的正态性检验通过。此外又对残差进行的独立性、同方差性、线性检验均通过，因此可以使用普通最小二乘来进行指数追踪。

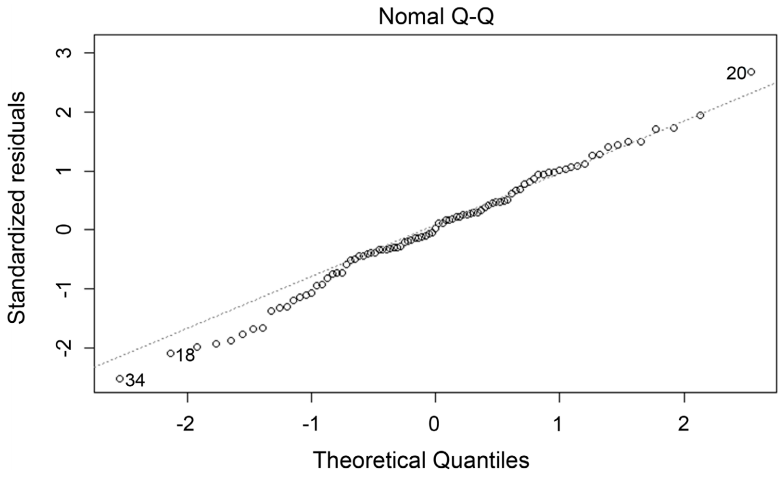

Figure 2. Lognormal quantile quantile plot
图 2. 对数正态 QQ 图

由图 3 可以看出，与单纯使用弹性网相比，拟合值和预测值都能更好地近似实际值，算出在训练集和测试集上的 RMSE 分别为 11.41、137.214，说明两步估计效果更好。

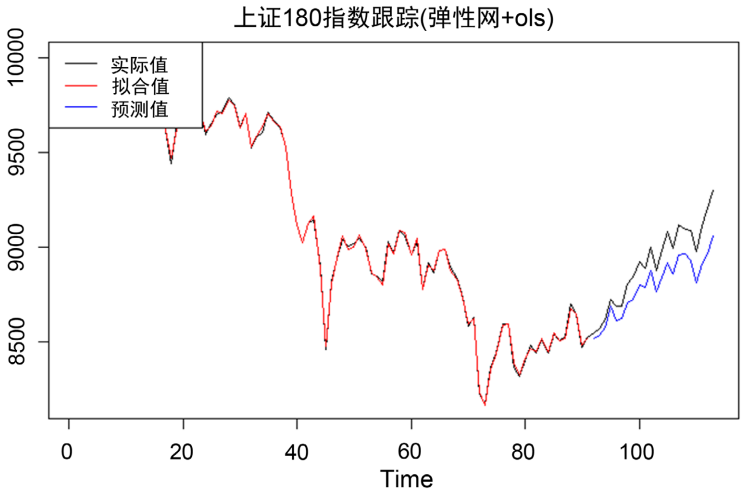

Figure 3. Diagram of the index tracking of elastic net + OLS
图 3. 弹性网 + 普通最小二乘指数追踪图

3.3. 弹性网 + 主成分回归

主成分回归(Principal Component Regression, PCR)是一种多元回归分析方法，它使用主成分分析(PCA)来处理自变量之间的多重共线性问题。在 PCR 中，首先对自变量进行主成分分析，然后选择前 k 个主成分来代替原始自变量，最后使用这些主成分来进行回归分析。PCR 的优点是可以处理多重共线性问题，同时可以减少自变量的数量，从而降低了模型的复杂度。它可以处理高维数据，减少自变量的数量，提高模型的预测精度和可解释性。

通过计算得到条件数为 1.130546 × 10^{18}，存在非常严重的多重共线性，因此本文使用主成分回归来解决这个问题。先对自变量进行主成分分析。

从图 4 可以看出，从第三个主成分开始，碎石图趋于一条平行于 x 轴的直线，并且考虑到前三个主成分的累计方差贡献率达到 88% (见表 1)，因此选用前三个主成分来进行主成分回归。

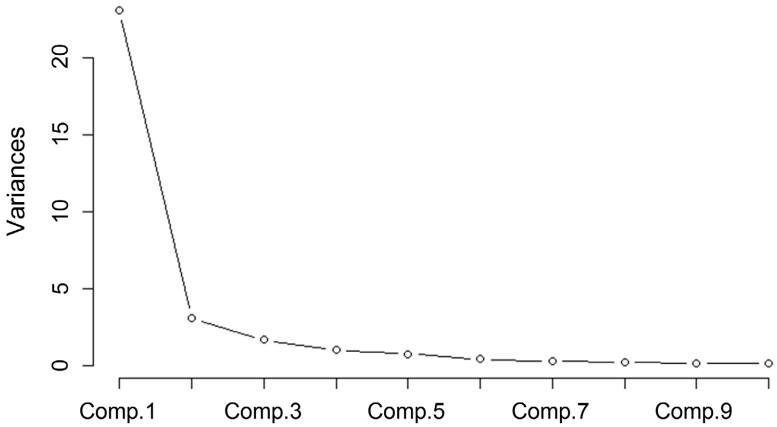

Figure 4. Scree plot
图 4. 碎石图

Table 1. Proportion of variance
表 1. 方差贡献率

	主成分 1	主成分 2	主成分 3
方差贡献率	73.07%	9.74%	5.31%
累计方差贡献率	73.07%	82.81%	88.12%

从图 5 可以看出，主成分回归较普通最小二乘回归效果更差了些，在训练集和测试集上的 RMSE 分别为 29.977、152.348，主成分回归虽然解决了多重共线性，但在残差平方和的角度却效果不太好，因此考虑更为稳健的分位数回归模型。

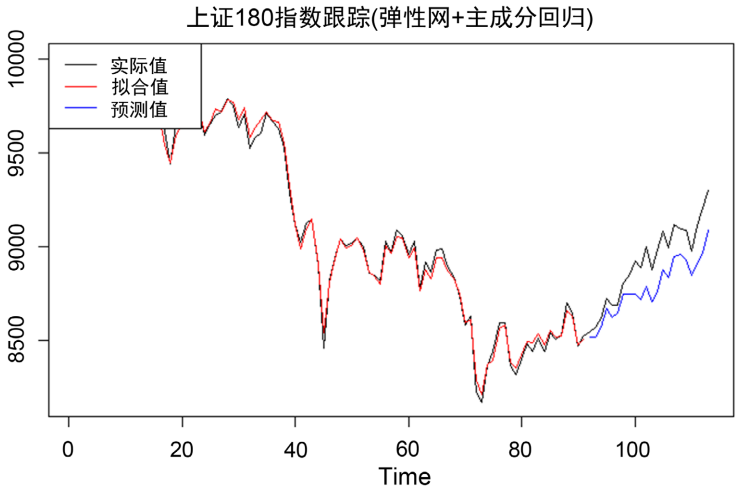

Figure 5. Diagram of the index tracking of elastic net + PCR
图 5. 弹性网 + 主成分回归指数追踪图

3.4. 弹性网 + 分位数回归

对于弹性网筛选出来的 32 个变量，利用分位数回归模型进行指数追踪。本文设置 0.05、0.25、0.5、0.75、0.95 五个分位点，并且进行分位数回归，最后绘制出五个指数追踪图。

从图 6 可以看出，5 个分位数回归模型在训练集上的表现都差不太多，但在测试集上 0.25 分位数回归能更好地近似指数的真实值，似乎也比前面所提方法都更好。最后算出各种模型在训练集和测试集上的均方根误差，从残差平方和的角度比较各种的表现。

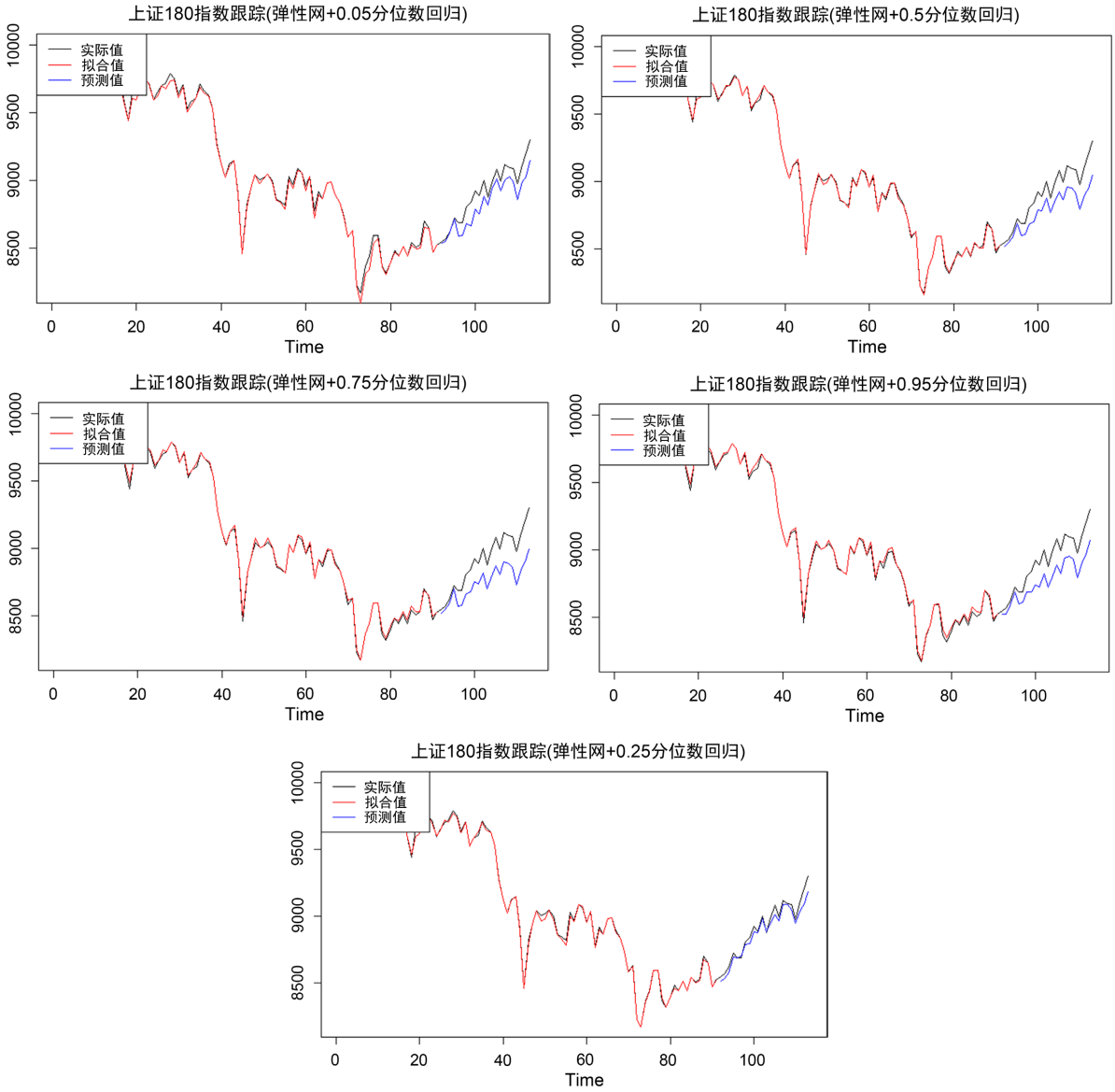

Figure 6. Diagram of the index tracking of elastic net + QR
图 6. 弹性网 + 分位数回归指数追踪图

由表 2 可知，弹性网+ols 训练集上的 RMSE 最小，说明这种方法的拟合表现最好，得到这个结果也不意外，原因是普通最小二乘估计的原理本身就是使得残差平方和达到最小。而在测试集上，几乎所有的两步估计都比单一的弹性网表现要好，而 0.25 分位数回归的 RMSE 最小的，0.05 分位数回归次之，可

以看出分位数回归在指数追踪中具有非常好的表现。

Table 2. RMSE of various methods
表 2. 各方法的均方根误差

方法	拟合残差平方和	拟合均方根误差 (RMSE)	预测残差平方和	预测均方根误差 (RMSE)
弹性网	32410.2	18.872	779227.9	188.201
弹性网 + ols	11846.9	11.410	414206.3	137.214
弹性网 + 主成分回归	81778.5	29.977	510621.1	152.348
弹性网 + 0.05 分位数	59376.7	25.544	245622.3	105.663
弹性网 + 0.25 分位数	21752.5	15.461	54887.7	49.949
弹性网 + 0.5 分位数	14315.7	12.543	462718.5	145.026
弹性网 + 0.75 分位数	23836.7	16.185	788559.6	189.324
弹性网 + 0.95 分位数	34821.7	19.562	537838.5	156.356

4. 结论

本文采用部分复制的策略来对上证 180 指数进行指数追踪，用均方根误差作为模型好坏的衡量标准。大量文献表明两步估计法较单一方法的追踪表现更好，故利用两步估计来追踪上证 180 指数，第一步使用弹性网来筛选股票，第二步考虑到模型的稳健性使用分位数估计来确定系数。在实证分析中与其他的一些方法作对比，结果显示两步估计方法优于单一方法，而在两步估计中分位数回归的指数追踪表现最佳，其中 0.25 分位数回归能更好的近似指数的真实值即追踪误差最小，追踪表现远远好于其他方法，因此可以使用 0.25 分位数回归去追踪上证 180 指数。

参考文献

[1] Kim, S. and Kim, S. (2020) Index Tracking through Deep Latent Representation Learning. *Quantitative Finance*, **20**, 639-652. https://doi.org/10.1080/14697688.2019.1683599

[2] Kwak, Y., Song, J. and Lee, H. (2021) Neural Network with Fixed Noise for Index-Tracking Portfolio Optimization. *Expert Systems with Applications*, **183**, 115298. https://doi.org/10.1016/j.eswa.2021.115298

[3] Bradrania, R., Pirayesh Neghab, D. and Shafizadeh, M. (2022) State-Dependent Stock Selection in Index Tracking: A Machine Learning Approach. *Financial Markets and Portfolio Management*, **36**, 1-28. https://doi.org/10.1007/s11408-021-00391-7

[4] Cao, Y., Li, H. and Yang, Y. (2022) Combining Random Forest and Multicollinearity Modeling for Index Tracking. *Communications in Statistics-Simulation and Computation*, 1-12. https://doi.org/10.1080/03610918.2022.2116050

[5] Wu, L., Yang, Y. and Liu, H. (2014) Nonnegative-Lasso and Application in Index Tracking. *Computational Statistics & Data Analysis*, **70**, 116-126. https://doi.org/10.1016/j.csda.2013.08.012

[6] Wu, L. and Yang, Y. (2014) Nonnegative Elastic Net and Application in Index Tracking. *Applied Mathematics and Computation*, **227**, 541-552. https://doi.org/10.1016/j.amc.2013.11.049

[7] Yang, Y. and Wu, L. (2016) Nonnegative Adaptive Lasso for Ultra-High Dimensional Regression Models and a Two-Stage Method Applied in Financial Modeling. *Journal of Statistical Planning and Inference*, **174**, 52-67. https://doi.org/10.1016/j.jspi.2016.01.011

[8] Li, N., Yang, H. and Yang, J. (2021) Nonnegative Estimation and Variable Selection via Adaptive Elastic-Net for High-Dimensional Data. *Communications in Statistics-Simulation and Computation*, **50**, 4263-4279. https://doi.org/10.1080/03610918.2019.1642484

[9] Li, N. and Yang, H. (2021) Nonnegative Estimation and Variable Selection under Minimax Concave Penalty for Sparse

High-Dimensional Linear Regression Models. *Statistical Papers*, **62**, 661-680. https://doi.org/10.1007/s00362-019-01107-w

[10] Chen, Q., Hu, Q., Yang, H., *et al*. (2022) A Kind of New Time-Weighted Nonnegative Lasso Index-Tracking Model and Its Application. *The North American Journal of Economics and Finance*, **59**, 101603. https://doi.org/10.1016/j.najef.2021.101603

[11] Li, N. (2020) Efficient Sparse Portfolios Based on Composite Quantile Regression for High-Dimensional Index Tracking. *Journal of Statistical Computation and Simulation*, **90**, 1466-1478. https://doi.org/10.1080/00949655.2020.1731750

[12] Li, N., Niu, Y. and Sun, J. (2022) Robust Sparse Portfolios for Index Tracking Based on M-Estimation. *Communications in Statistics-Simulation and Computation*, 1-13. https://doi.org/10.1080/03610918.2022.2112054