Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5151-5158
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135517

Hans汉斯

结合聚类和时间加权非负岭回归指数追踪模型

练桂伶

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2023年8月12日；录用日期：2023年10月11日；发布日期：2023年10月20日

摘　要

指数追踪是一种特殊的被动投资管理形式，其目的在于从一个目标指数包含的所有成分股中选择一部分成分股来建立股票投资组合，用其来追踪目标指数，尽可能的得到一个接近股票市场的累计回报。本文将变量聚类与经典统计模型岭回归相结合，提出聚类和时间加权非负岭回归指数追踪模型，该模型首先使用变量聚类法选择部分股票构建投资组合，然后考虑到时间因素对指数追踪的影响，以及股票市场不允许卖空约束，在岭回归的基础上引入了时间加权函数和对权重的非负约束。为了验证所提出模型的性能，本文将聚类和时间加权非负岭回归指数追踪模型与现有模型进行比较。实验结果表明，使用聚类和时间加权非负岭回归指数追踪模型可以得到较小的追踪误差，所构造的投资组合具有较好的追踪效果。

关键词

指数追踪，变量聚类，时间加权非负岭回归，乘法更新

Combining Clustering and Time-Weighted Non-Negative Ridge Regression Index Tracking Models

Guiling Lian

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Aug. 12th, 2023; accepted: Oct. 11th, 2023; published: Oct. 20th, 2023

Abstract

Index tracking is a special form of passive investment management, which aims to select a subset of all constituents of a target index to build a stock portfolio, and use it to track the target index to obtain a cumulative return as close to the stock market as possible. This paper combines variable clustering with the classical statistical model ridge regression, and proposes a clustering and

文章引用: 练桂伶. 结合聚类和时间加权非负岭回归指数追踪模型[J]. 运筹与模糊学, 2023, 13(5): 5151-5158.
DOI: 10.12677/orf.2023.135517

time-weighted non-negative ridge regression index tracking model, which first uses the variable clustering method to select some stocks to construct a portfolio, and then considers the impact of time factors on index tracking and the stock market does not allow short selling constraints, and introduces a time-weighted function and a non-negative constraint on weights on the basis of ridge regression. To verify the performance of the proposed model, this paper compares the clustering and time-weighted non-negative ridge regression exponential tracking model with the existing model. Experimental results show that the tracking error can be obtained by using clustering and time-weighted non-negative ridge regression index tracking model, and the constructed portfolio has better tracking effect.

Keywords

Exponential Tracking, Variable Clustering, Time-Weighted Non-Negative Ridge Regression, Multiplication Update

1. 引言

指数追踪是一种特殊的被动投资管理形式，其目的在于从一个目标指数包含的所有成分股中选择一部分成分股来建立一个股票投资组合，用其来追踪目标指数，尽可能的得到一个接近股票市场的累计回报。指数追踪策略主要有两种方法，分别为完全复制和部分复制。完全复制是最直接简单的方法，其根据目标指数包含的所有成分股的权重来投资于所有的成分股。然而，这种指数追踪策略往往会产生更显著的交易成本和更高的管理费用。而部分复制是更多作者的选择，其只需要选择目标指数的一部分成分股来进行指数追踪，然后最小化追踪误差。虽然部分复制方法会产生一些追踪误差，但是其交易成本和管理成本都相对较低，因此更多的投资者更倾向于选择使得追踪误差较小的部分复制方法来进行指数追踪。现有的指数追踪方法主要有基于传统统计方法、基于启发式的方法和基于学习的方法。Wu [1]提出了线性整数规划模型，使用基准投资组合的成分股来追踪目标投资组合，且使用拉格朗日和半拉格朗日方法来计算模型的最优解。Chen and Kwon [2]提出了一个 0~1 整数程序模型用来提高所选的投资组合与目标指数之间的相似性。Yang 和 Wu [3]提出了非负自适应 LASSO 方法进行变量选择来解决指数追踪问题。Guastaroba [4]提出了一个具有内核搜索的启发式模型来解决指数追踪问题。该模型包括对投资组合更新时产生总交易成本进行约束，并且证明了该模型具有有效性和高效性。Mutunge 和 Hangland [5]设计了一个启发式算法专注解决指数追踪问题，使用二次函数指数的协方差矩阵返回系数矩阵。Fu [6]提出了一种基于监督学习的堆叠股票选择模型，采用遗传算法选择股票特征，并根据回归波动率排序对股票进行标记。最近，随着深度学习在人工智能领域的出现，深度学习技术开始被广泛应用于指数追踪中。Ouyang [7]等人使用了深度学习算法来追踪股票的表现，即使用深度自编码器选择股票数组成投资组合，然后使用深度神经网络动态确定每只股票的权重。Lu [8]使用长短期记忆(LSTM)实现了一个具有强化学习作为代理的稳健性和可行性的学习模型。从以往研究文献可以看出，基于统计方法是最成熟的方法，但是这种方法有一定的局限性，其需要大量的计算，并且当数据集的协方差矩阵为病态或非正态时，该方法就及其不稳定。基于启发式算法在选择投资组合方面是很有效的，但是针对高维数据集的指数追踪问题时效率就十分低下，容易落入局部最小值，往往导致次优的投资组合。基于学习的方法能有效的从股价中

提取特征和有效信息来预测其走势，但基于学习的方法对股票市场的稳定性很敏感，往往有较大的波动。

为了解决以上问题，本文提出了一个结合聚类和时间加权非负岭回归指数跟踪模型，旨在从基准股票中选择部分股票组成投资组合，并确定每个组成股票的投资权重。一方面，本文利用变量聚类法选择股票，因为其可以有效地从股票价格中提取特征和有效信息，并且在处理大数据集时相对可伸缩和高效率且时间复杂度低。另一方面，使用时间加权非负岭回归确定投资组合的权重，与其他回归方法相比，岭回归在解决多重共性方面具有更好的效果。最后在不同数据集上应用该模型与其他模型做比较，研究结果表明该模型有更好的追踪效果。

2. 变量聚类

聚类分析可以看作将数据集中相似性高的变量聚成一类，换句话说就是对收集的数据进行无标签的分类，聚类分析在很多学科包括统计学、生物学、数学等应用广泛，在广泛学者的研究中，聚类技术创新出了很多的方法，这些创新方法的共同点都是用不同的方法衡量数据之间的相似性，将变量分到不同的类中去。

近年来，有很多学者使用变量聚类法进行降维，并取得不错的结果，变量聚类可以从数据中提取特征或有效信息，并且在处理大数据集时相对可伸缩和高效率且时间复杂度低，因此本文采用 Kmeans 聚类算法对股票进行降维，选择股票构建投资组合。

本文将基准指数中的股票聚为 K 个簇，计算每个簇中所有股票到中心的距离，并选择距离中心位置小于总距离的 10%的股票进入投资组合中。具体来说，本文选股的详细过程如算法 1 所示。

Step 1：每个特征进行最大值最小值归一化。

Step 2：根据肘部法则，确定聚类数 k。

Step 3：对于每一个类，计算所有股票到类中心得距离总和 S_i 以及计算离类中心最近的 d 只股票的距离和 s_i。

Step 4：选择满足 $\frac{s_i}{S_i} \geq 10\%$ 的 d 只股票，加入候选股票中。

Step 5：最后汇总每个类中选择的股票组成投资组合共 K 只股票。

3. 时间加权非负岭回归

3.1. 岭回归

处理线性回归的方法有很多，最基本方法是最小二乘法，由于股票数据存在多重共线性，最小二乘并不适用于对股票数据的回归系数估计，岭回归是一种专门用于处理多重共线性数据的有偏估计方法，岭回归实际上是一种改良的最小二乘，它舍弃了最小二乘的无偏性和部分精确度来寻求效果稍微较差但是更符合实际分析的回归过程。岭回归的定义为：

$$\hat{\beta}(k)=\left(X'X+kI\right)^{-1}X'y$$

其中，X 是标准化后的矩阵，$\hat{\beta}(k)$ 为岭回归的估计，适当的确定 K 值，可以解决最小二乘在求系数的向量时矩阵无法求逆的问题，从而减少多重共线性的影响。因此当 $K=0$ 时，岭回归的是最小二乘。

岭回归的损失函数定义如下：

$$J(w)=\left\|Y_t-\hat{Y}_t\right\|_2^2+\lambda_n\sum_{t=1}^{T}\left\|w\right\|_2^2$$

假设 $x=\left(x_1,\cdots,x_N\right)^{\mathrm{T}}$，则 $\left\|x\right\|_2^2=\sum_{i=1}^{N}x^2$。

3.2. 时间加权非负岭回归

由于金融时间序列数据的不稳定和高噪声，数据越接近当前时间，对未来的影响就越大，提供的信息就越多，因此将各时间段的权重逐渐增大是一种有效的时间序列处理方法。本文引用了 Cao 和 Tay [9] (2000)提出的时间加权函数如下所示：

$$\lambda_t = \frac{2}{1+\mathrm{e}^{\alpha-\frac{2\alpha t}{T}}},\ t=1,2,\cdots,T-1 \tag{1}$$

α 是调整参数或贴现率，Tay 等人认为，$\alpha=1$ 更适合用于财务数据分析，因此本研究采用了 $\alpha=1$。

结合股票市场不允许卖空的市场约束下，提出了时间加权非负岭回归指数追踪模型：

$$\begin{cases} \underset{w\geq 0}{\hat{w}(\lambda_n)} = \arg\min \lambda_t \left\| Y_t - \hat{Y}_t \right\|_2^2 + \lambda_n \|w\|_2^2 \\ s.t \sum_{i=1}^{N} w_i = 1 \ \ w_i > 0 \ \ i=1,2,\cdots,N \end{cases} \tag{2}$$

其中 λ_n 为正则化参数，N 代表的是投资组合的股票数。该模型为二次规划问题。本文将使用乘法更新算法求解该二次规划问题，求解步骤如下：

$$\begin{cases} \text{minimize} & F(v) = \frac{1}{2} v'Av + b'v, \\ \text{subject} & v \geq 0, \end{cases}$$

其中 $\upsilon = (\upsilon_1, \upsilon_2, \cdots, \upsilon_n)'$ 是一个 n 维列向量，$A \in \mathbb{R}^{n\times n}$ 是一个半正定矩阵。

$$A_{ij}^{+} = \begin{cases} A_{ij} & \text{if } A_{ij} > 0, \\ 0 & \text{otherwise}, \end{cases} \text{ and } A_{ij}^{-} = \begin{cases} \left|A_{ij}\right| & \text{if } A_{ij} < 0, \\ 0 & \text{otherwise}, \end{cases}$$

其中 $F_a(v) = \frac{1}{2} v'A^{+}v$, $F_b(v) = b'v$, $F_c(v) = \frac{1}{2} v'A^{-}v$, $a_i = \frac{\partial F_a}{\partial v_i} = \left(A^{+}v\right)_i$, $c_i = \frac{\partial F_c}{\partial v_i} = \left(A^{-}v\right)_i$

则迭代步骤为：

$$\left[\frac{-b_i + \left(b_i^2 + 4a_i c_i\right)^{\frac{1}{2}}}{2a_i} \right] \cdot v_i^{(m)} \to v^{(m+1)_i}$$

因此由乘法更新算法求解时间加权非负岭回归模型可以化为如下：

$$\begin{aligned} F(w) &= \lambda_t \left\| Y_t - \hat{Y}_t \right\|_2^2 + \lambda_n \sum_{t=1}^{T} \|w\|_2^2 \\ &= (Y - xw)^{\mathrm{T}} \lambda_t (Y - xw) + w^{\mathrm{T}} \lambda_n w \\ &= w^{\mathrm{T}} x^{\mathrm{T}} \lambda_t x w - 2Y\lambda_t x W + Y^{\mathrm{T}} \lambda_t Y + w^{\mathrm{T}} \lambda_n w \\ &= \frac{1}{2} w^{\mathrm{T}} 2\left(x^{\mathrm{T}} \lambda_t x + \lambda_n 1'\right) w + \left(-2x^{\mathrm{T}} \lambda_t Y\right)^{\mathrm{T}} w + Y^{\mathrm{T}} \lambda_t Y \end{aligned} \tag{3}$$

从以上等式可以看出对于时间加权非负岭估计的二次规划问题解的公式为：

$$\begin{aligned} &\text{minmiz } F(w) = \frac{1}{2} w^{\mathrm{T}} 2\left(x^{\mathrm{T}} \lambda_t x + \lambda_n 1'\right) w + \left(-2x^{\mathrm{T}} \lambda_t Y\right)^{\mathrm{T}} w + Y^{\mathrm{T}} \lambda_t Y \\ &s.t.\ w \geq 0 \end{aligned} \tag{4}$$

其中， $A=2\left(x^{\mathrm{T}}\lambda_t x+\lambda_n 1'\right)$， $b=-2x^{\mathrm{T}}\lambda_t Y$ 利用乘法更新算法可以得出时间加权非负岭估计量。下面将结合广义交叉验证法则确定正则化参数的 λ_n 值，具体步骤如下：

Step 1：设置岭参数 λ_n 值： $\lambda_1,\lambda_2,\cdots,\lambda_n$ ，代入： $F\left(w\right)$

Step 2：用乘法更新算法解： $w_1,w_2,\cdots,w_n$ ，

Step 3：代入 $GCV\left(w\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i-\widehat{R_i\left(w\right)}}{1-\frac{1}{n}tr\left(L\left(w\right)\right)}\right)^2$ ，得到 $GCV_1,GCV_2,\cdots,GCV_n$ ，使得 GCV 最小时的 w,k 值，其中 $L\left(w\right)=x\left(x^{\mathrm{T}}\lambda_t x+\lambda_n I\right)^{-1}x^{\mathrm{T}}\lambda_t$ 。

4. 实证分析

4.1. 数据收集

本文共选择了三个数据集，分别为代表稳定金融市场的标准普尔 500 指数(美国市场)，由 503 只股票加指数价格系列组成。代表新兴市场的沪深 300 指数(中国市场)由 300 只股票组成，恒生指数(中国市场)由 76 只股票组成。本文的所有数据集均来自于 Choice 金融终端，数据集收集自 2018 年 12 月 3 日至 2023 年 2 月 16 日的收盘价数据，然后通过收盘价数据计算收益率数据，收益率的公式如下：

$$r_t=\frac{p_t}{p_{t-1}}-1,\ t=1,\cdots,T \tag{5}$$

r_t 表示股票在第 t 日的收益率， p_t 为股票在 t 日的回报率，由于有个别股票存在某个时间段数据缺失的情况，本文将数据缺失的股票全部删除，得到每个数据集最终包含的股票数。然后根据变量聚类法可以得到每个数据集最终确定的投资组合的成分股数见表 1 所示。

Table 1. The number of base and portfolio stocks for the dataset
表 1. 数据集的基准股票数和投资组合股票数

指数	时间	基准股票数	投资组合股票数
标准普尔 500	2018/12/03~2023/02/16	489	55
沪深 300	2018/12/03~2023/02/16	255	40
恒生指数	2018/12/03~2023/02/16	67	20

指数追踪的目标是在基数约束下选择资产及其权重，从而使得所获得的追踪回报接近指数的返回。这种接近度的度量由追踪误差决定，如公式(6)所示。

$$MSE=\frac{1}{T}\sum_{t=1}^{T}\left(Y_t-\hat{Y}_t\right)^2 \tag{6}$$

因此本文将使用追踪误差 MSE 来判断模型的追踪能力。

4.2. 结果分析

首先，将三个数据集的日收益率数据集分为训练集和测试集，其中前 80%为训练集，后 20%为测试集，然后编写 R 代码，训练聚类 + 时间加权非负岭回归指数追踪模型。为了说明聚类 + 时间加权非负指数追踪模型的追踪效果，本文采用了相同的数据训练方式对 Chen [10]在(2022)年提出的时间加权非负 LASSO 指数追踪模型进行了训练，它们在测试集上的追踪误差结果见表 2 所示。

Table 2. Tracking error results
表 2. 追踪误差结果

指数	模型	股票数	MSE
标准普尔 500	聚类 + 时间加权非负岭估计	55	4.2522e−05
	时间加权非负 LASSO	55	5.4892e−05
沪深 300	聚类 + 时间加权非负岭估计	40	1.0534e−05
	时间加权非负 LASSO	40	5.5843e−05
恒生指数	聚类 + 时间加权非负岭估计	20	3.9214e−05
	时间加权非负 LASSO	20	7.6612e−05

从表中可以得出，第一步使用变量聚类降维得到标准普尔500指数最终选择55只股票进入投资组合中，沪深300指数和恒生指数分别选择40和20只股票进入投资组合中，为了便于比较本文调节时间加权非负LASSO 的参数 λ_n 的值，使得两个模型在最终构建的投资组合的股票数相等。在最终的指数追踪结果中，可以看出在三个数据集中，本文提出的模型聚类 + 时间加权非负岭回归指数追踪的MSE在测试集小于时间加权非负LASSO，因此本文提出的模型有较小的追踪误差，所构造的投资组合具有较好的追踪效果。

图1到图3比较了结合聚类和时间加权非负岭回归模型和时间加权非负LASSO在三个数据集上指数追踪图。从图中可以明显看出，对标准普尔 500 数据，结合聚类和时间加权非负岭回归明显优于时间加权非负LASSO；对沪深300数据和恒生指数数据聚类和时间加权非负岭回归同样明显优于时间加权非负LASSO。因此从指数追踪图也可以得出与追踪误差结果相同的结论。

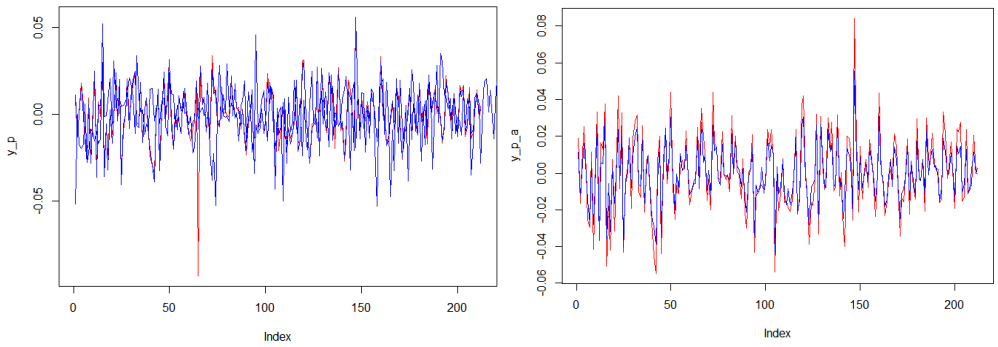

Figure 1. Data: Standard & Poor's 500, the left figure is the clustering + time-weighted non-negative ridge regression model results, and the right figure is the time-weighted non-negative LASSO model results
图 1. 数据：标准普尔 500，左图是聚类 + 时间加权非负岭回归模型结果，右图是时间加权非负 LASSO 模型结果

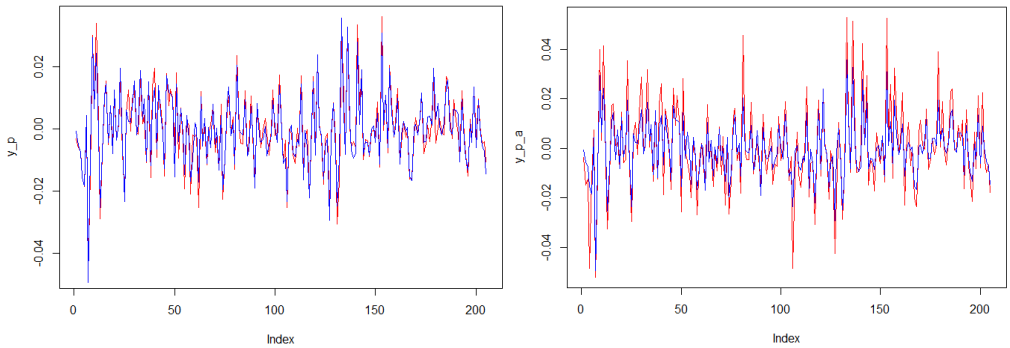

Figure 2. Data: CSI 300, the left figure is the result of clustering + time-weighted non-negative ridge regression model, and the right figure is the result of the time-weighted non-negative LASSO model
图 2. 数据：沪深 300，左图是聚类 + 时间加权非负岭回归模型结果，右图是时间加权非负 LASSO 模型结果

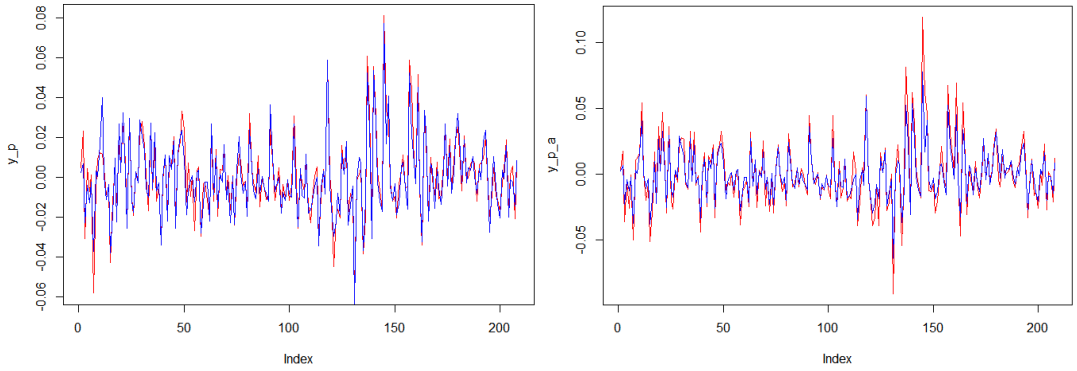

Figure 3. Data: Hang Seng, the left figure is the result of clustering + time-weighted non-negative Ridge regression model, and the right figure is the result of the time-weighted non-negative LASSO model
图 3. 数据：恒生指数，左图是聚类 + 时间加权非负岭回归模型结果，右图是时间加权非负 LASSO 模型结果

综上所述，从指数追踪图都很明显可以看出本文提出的聚类 + 时间加权非负岭回归模型明显优于时间加权非负 LASSO。

5. 结论

本文研究了三个数据集分别是标准普尔 500、沪深 300、恒生指数，维数从 489、255 到 67 维，对于高维数据本文首先对数据进行降维，再进行指数追踪。通过变量聚类法对股票数据进行降维，选择成分股构建投资组合，其中使用了肘部法则确定聚类数，再从每个类中选择股票。由于股票数据存在多重共线性问题，因此本文使用了无偏估计岭回归进行分析。然后考虑时间对金融数据的影响以及股票市场不允许卖空的约束构建了时间加权非负岭回归模型。在求解参数时，使用了广义交叉验证求解岭参数，以及乘法更新算法计算岭估计。

本文得到的结论如下：使用变量聚类进行降维，标准普尔 500 数据保留了 55 只股票，沪深 300 数据保留了 40 只股票，恒生数据保留了 20 只股票。且三个数据集中本文提出的模型聚类 + 时间加权非负岭回归指数追踪的 MSE 小于时间加权非负 LASSO，因此本文提出的模型有较小的追踪误差，所构造的投资组合具有较好的追踪效果。

综合以上研究，本文提出的方法在实际追踪中误差表现最好，并且有较好的预测性和稳定性，构建的投资组合稀疏性达到了较优，符合股票市场的要求，保证了非空投资。并且本文将机器学习算法与统计方法相结合，汲取各自的优点运用到股票数据中去。因此本文的研究方法和实证分析对指数型投资者具有重要的参考价值。

参考文献

[1] Wu, K. and Costa, G. (2017) A Constrained Cluster-Based Approach for Tracking the S&P 500 Index. *International Journal of Production Economics*, **193**, 222-243. https://doi.org/10.1016/j.ijpe.2017.07.018

[2] Chen, C. and Kwon, R.H. (2012) Robust Portfolio Selection for Index Tracking. *Computers & Operations Research*, **39**, 829-837. https://doi.org/10.1016/j.cor.2010.08.019

[3] 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

[4] Mutunge, P. and Haugland, D. (2018) Minimizing the Tracking Error of Cardinality Constrained Portfolios. *Computers & Operations Research*, **90**, 33-41. https://doi.org/10.1016/j.cor.2017.09.002

[5] Scozzari, A., Tardella, F., Paterlini, S. and Krink, T. (2013) Exact and Heuristic Approaches for the Index Tracking Problem with UCITS Constraints. *Annals of Operations Research*, **205**, 235-250.

https://doi.org/10.1007/s10479-012-1207-1

[6] Fu, X.Y., Du, J.H., Guo, Y.F., Liu, M.W., Dong, T., and Duan, X.W. (2018) A Machine Learning Framework for Stock Selection. Quantitative Finance, Portfolio Management. https://arxiv.org/abs/1806.01743

[7] Ouyang, H., Zhang, X. and Yan, H. (2019) Index Tracking Based on Deep Neural Network. *Cognitive Systems Research*, **57**, 107-114. https://doi.org/10.1016/j.cogsys.2018.10.022

[8] Lu, W. (2017) Agent Inspired Trading Using Recurrent Reinforcement Learning and LSTM Neural Networks. Quantitative Finance, Computational Finance. https://arxiv.org/abs/1707.07338

[9] Cao, L.J. and Tay, F. (2000) Feature Selection for Support Vector Machines in Financial Time Series Forecasting. *International Conference on Intelligent Data Engineering & Automated Learning—IDEAL* 2000. *Data Mining, Financial Engineering, and Intelligent Agents*, Springer-Verlag, 268-273. https://doi.org/10.1007/3-540-44491-2_38

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