Finance 金融, 2019, 9(4), 327-340
Published Online July 2019 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2019.94041

Hans汉斯

Research on Multi-Factor Stock Selection Model Based on EKPCA Algorithm

Yu Huang, Houqing Fang, Lingfei Dai, Tingting Chen

Faculty of Science, Jiangsu University, Zhenjiang Jiangsu
Email: doudou@ujs.edu.cn

Received: Jun. 20th, 2019; accepted: Jul. 3rd, 2019; published: Jul. 12th, 2019

Abstract

The multi-factor stock selection model is the mainstream method in quantitative investment. This paper introduces the Efficient Kernel Principal Component Analysis (EKPCA) algorithm for the first time. The high-efficiency kernel principal component is used as the independent variable to establish the regression equation to predict the rate of return and construct a multi-factor stock selection model. Based on the empirical analysis of the constituents of SSE 180, this paper selects more than 50 impact factors including fundamentals, technical indicators and investor sentiment indicators, and uses the EKPCA algorithm to determine the basic model and extracts high-efficiency kernel principal components in the high-dimensional feature space. Compared with the classical KPCA algorithm, the EKPCA algorithm has higher feature extraction efficiency. The backtest results show that the beta coefficient and Sharpe ratio of the constructed portfolio are better than the market benchmark level in the selected time period, which indicates that the model has a better stock picking effect.

Keywords

EKPCA Algorithm, Multi-Factor Stock Selection, The General Entropy, Feature Extraction, Kernel Function

基于EKPCA算法的多因子选股模型研究

黄　钰，房厚庆，戴凌飞，陈婷婷

江苏大学理学院，江苏　镇江
Email: doudou@ujs.edu.cn

收稿日期：2019年6月20日；录用日期：2019年7月3日；发布日期：2019年7月12日

摘　要

多因子选股模型是量化投资中的主流方法。本文首次引入高效的核主成分分析(Efficient Kernel Principal

文章引用: 黄钰, 房厚庆, 戴凌飞, 陈婷婷. 基于 EKPCA 算法的多因子选股模型研究[J]. 金融, 2019, 9(4): 327-340.
DOI: 10.12677/fin.2019.94041

Component Analysis, EKPCA)算法，以高效的核主成分为自变量建立回归方程预测收益率，构建多因子选股模型。本文基于上证180的成分股进行实证分析，选取包含基本面、技术指标及投资者情绪指标等50多个影响因子，引用EKPCA算法确定基本模式，在高维特征空间提取高效核主成分。与经典KPCA算法对比，EKPCA算法具有更高的特征抽取效率。回测结果显示，构造的投资组合的贝塔系数和夏普比率在所选时间段内均优于市场基准水平，这表明该模型具有较好的选股效果。

关键词

EKPCA算法，多因子选股，通用熵，特征提取，核函数

1. 引言

人工智能技术的高速发展，机器学习算法的持续改进，使得量化投资更具实用价值。1971 年，美国巴克莱国际投资管理公司发行了世界上第一只指数基金，这标志着量化投资交易正式拉开帷幕。在美国，量化投资交易已有将近 50 年的历史，量化投资交易量从开始到现在已经占到美国金融市场交易量的 30%以上。量化投资起源于美国，却在全球资本市场的浪潮中得到快速的发展。国外成熟金融市场的发展经验表明，利用量化投资可以有效提高金融交易市场的资金流动性。在 2004 年，国内首次出现量化基金——光大保德信量化核心基金，此后因为各种原因，发展比较缓慢。最主要的量化投资策略就是多因子策略、CTA 策略和对冲策略。其中，多因子策略 2007 年在国内出现，此后一直占据量化投资策略的主导地位；CTA 策略在 2011 年进入国内市场后，并没有较大发展空间，直到 2015 年才开始发展起来；对冲策略出现在 2012 年左右，主要依托股指期货的对冲效应[1]。量化投资减少了主观判断等因素的影响，比传统投资更具科学性，越来越受投资者的青睐。

国际上经典的选股模型是资本资产定价模型(CAPM)、套利定价理论模型(APT)、法玛和弗伦奇的三因子模型。当前，中国的股票市场对于多因子模型的研究还处在一个初级阶段，绝大多数的研究都局限于证券公司的专业研究部门，在学术界算是比较新兴的领域。

国内学者通过借鉴国外先进的量化模型，构建了不同类型的量化选股模型。范烨[2]构建了多因子模型进行选股研究，对多因子进行排序打分，并且确定相关阈值对因子进行筛选；李娜[3]等提出基于 k-means 聚类的选股模型，该模型利用数据的自动学习来对数据进行特征分类选取因子；苏治[4]构建了基于核主成分遗传算法改进的支持向量回归机人工智能选股模型，提高了传统模型的预测精度；贾秀娟[5]使用随机森林处理变量，提出基于随机森林的支持向量机模型进行选股研究，提高了模型的识别精度。其中，多因子选股模型有两种判别方法，一种是打分法[2] [6]，另一种是回归法[7]。与打分法相比，回归法可以根据股票市场上的突发情况，能够及时地调整模型对各个因子的敏感度，而且简单、快速，更便于程序化交易。因此，本文采用回归法对多因子模型进行实证研究。

我国股票市场存在 T + 1 制度，缺乏有效的做空机制和金融衍生工具，这些都限制了量化投资在中国市场的发展。我国 A 股市场发展还不完善，风格多变，多处于弱有效状态，因子之间关系复杂，内在混乱，各个财务指标之间的线性关系不明显，这里利用普通的线性降维效果并不理想。为了探究适应我国 A 股市场的量化投资策略，本文提出高效的核主成分分析方法，在高维空间中提取主要的非线性特征。

经典 KPCA 方法常用于特征提取、图像识别、人脸识别等[8]。KPCA 作为抽取复杂数据特征的一种非线性方法，当训练样本的规模很大时，它的抽取效率并不高，因此本文提出一种新型 EKPCA 算法来提高特征抽取效率。EKPCA 算法的基本思想是剔除对特征抽取贡献小的训练样本，确定贡献大的训练样本，即确定基本模式，运用特征相关分析来逐个确定基本模式，基本模式的个数可由用户根据需要自己设置。再用已确定的基本模式重新建立 KPCA 模型，在高维空间进行特征抽取，将抽取的主成分进行收益率回归。本文首次将该算法应用到多因子选股模型中，为量化投资策略提供新的研究思路。

2. 影响因子选取

本文选取的因子包括上市公司的基本面数据，技术指标数据及投资者情绪指标数据[9]等 50 多个因子。表 1 给出了所选因子：

Table 1. Selected impact factors
表 1. 选取的影响因子

因子类型	因子名称
规模类因子	流通 A 股(十亿股)、总股本(十亿股)、总市值(十亿)、股东权益
估值类因子	市盈率(PE)、市净率(PB)、市销率(PS)、市现率(PCF)、每股净资产(BPS)、股息率
风险类因子	Alpha、Beta、Sharpe、Treynor、Jensen、可决系数R^2
每股指标因子	每股收益(EPS)、每股营业收入、每股息税折旧摊销前利润
盈利能力因子	净资产收益率(ROE)、总资产收益率(ROA)、销售净利率、投入资本回报率(ROIC)、销售毛利率、净资产/营业总收入、息税前利润/营业总收入
收益质量因子	经营活动净收益/利润总额、营业利润/利润总额
资本结构因子	资产负债率、权益乘数
短期偿债能力因子	流动比率、速动比率、现金比率
营运能力因子	总资产周转率、营业周期
成长能力因子	每股收益同比增长率、净利润同比增长率、营业利润同比增长率、现金流量净额同比增长率、营业收入同比增长率、净资产同比增长率
技术指标因子	成交量、换手率、MACD 指数平滑异同平均、DMI 趋向指标、RSI 相对强弱指标、ROC 变动速率、KDJ 随机指标、OBV 能量潮、STD 标准差
分析师预测因子	预测 PEG、预测净利润增长率、预测营业收入增长率
情绪能量因子	PSY 情绪指标、BRAR 人气意愿指标

3. 多重共线性检验

本文选取的影响因子比较多，因子之间可能存在较高的相关性，这会严重影响选股模型的准确性，使得之后的回归预测失效。因此，本文引用方差膨胀因子(VIF)检验法[10]，影响因子 x_i 的方差膨胀因子记为 VIF，它的计算方法为：

$$\mathrm{VIF}=\frac{1}{1-R^2}$$

式中，R^2 是以 x_i 为因变量时对其它影响因子 x_j，$j\neq i$ 回归的模型拟合优度。如果最大的 VIF 超过 10，常常表示多重相关性将严重影响回归参数的估计值。基于我国股票市场弱有效且经济因素复杂的背景下，多维度的因子之间不存在明显的线性相关关系。因此，本文引用 EKPCA 算法进行非线性降维解决上述问题。

4. 基于 EKPCA 算法的多因子选股模型建立

4.1. EKPCA 算法的基本原理

经典 KPCA 是传统 PCA 的一种非线性拓展，其目的是从数据中抽取最具表现力的特征[11]。因子之间存在非线性相关关系，不能直接用 PCA 进行降维，而 KPCA 的基本思想是将原始训练样本通过一个合适的非线性隐函数 $\Phi(x)$ 映射到高维特征空间中，原空间中非线性可分变量很大程度可以转化为高维特征空间中的线性可分变量，然后在高维空间中运行 PCA，从而抽取数据的特征。通过研究映射到高维空间中的点的几何性质，我们知道点之间的距离和夹角都可以用核函数 κ 来表示，不需要知道 $\Phi(x)$ 的具体形式。常用的核函数有线性核、多项式核、高斯核、拉普拉斯核、Sigmoid 核等，本文根据模型需要主要引入表 2 中的两种核函数：

Table 2. Common kernel functions
表 2. 常见的核函数

名称	表达式	参数
多项式核	$\kappa(x_i,x_j)=(x_i^{\mathrm{T}}x_j+c)^d$	$c\geq 0$，d 为多项式的次数且是正整数
高斯核	$\kappa(x_i,x_j)=\exp\left(-\frac{\lVert x_i-x_j\rVert^2}{2\sigma^2}\right)$	$\sigma>0$ 为高斯核的带宽(width)

其中，x_i，x_j 为原始输入空间的训练样本。利用核函数建立核矩阵 K，核矩阵 K 其实是内积矩阵，且一定是半正定矩阵[12]，其中

$$K_{ij}=\left\langle\Phi(x_i),\Phi(x_j)\right\rangle=\Phi(x_i)^{\mathrm{T}}\Phi(x_j)=\kappa(x_i,x_j),\ i=1,2,\cdots,N;\ j=1,2,\cdots,N \tag{1}$$

式中，x_i，x_j 为训练样本，N 为训练样本的个数。

由上述表达式可知核矩阵的规模与训练样本的个数成正比，则 KPCA 的计算效率与训练样本的个数成反比，多因子选股模型一般选取的股票数量都较多，KPCA 抽取效率较低。近年来，很多算法被提出用来提高 KPCA 的计算效率[13]。本文引用其中一种特征抽取效果较好的 EKPCA 算法，该算法可以用小部分训练集(基本模式)的线性组合来近似表达 KPCA 的特征抽取效果。对全体训练样本进行特征抽取时，每个特征分量其实就是一系列核函数的线性组合，训练样本越多，特征抽取效率越低。同时，在特征抽取过程中，有些训练样本对特征抽取的贡献非常小。EKPCA 算法的基本思想是在进行特征抽取前，先剔除对特征抽取贡献小的训练样本，保留贡献大的训练样本，保留下来的训练样本即基本模式，这里运用特征相关分析来逐个确定基本模式，基本模式的个数由用户自己设置。再重新构建 KPCA 模型。

EKPCA 算法与 KPCA 算法相比主要有两个优点：第一，基本模式的个数一般远远小于全体训练样本的个数，这使得 EKPCA 算法比 KPCA 算法的抽取效率高；第二，在特征抽取时，KPCA 需要对 $N\times N$ (N 是全体训练样本的个数)的核矩阵进行特征分解，而 EKPCA 只需要对 $s\times s$ (s 是基本模式的个数)的核矩阵进行特征分解，这说明 EKPCA 比 KPCA 需要更少的存储空间[14]。

4.2. 核函数参数选择

首先，对训练样本进行 z-score 标准化处理，即使处理后的数据均值为 0，标准差为 1。

其次，引入最大通用熵来确定多项式核函数的参数 c 和 d 以及高斯核函数的参数 θ。主要步骤如下：

步骤 1：定义特征空间中点之间的余弦距离

$$c_{ij}=\frac{\Phi\left(x_i\right)^{\mathrm{T}}\Phi\left(x_j\right)}{\left\|\Phi\left(x_i\right)\right\|\cdot\left\|\Phi\left(x_j\right)\right\|},\ i=1,2,\cdots,N;\ j=1,2,\cdots,N \tag{2}$$

将(1)式带入(2)式得到下式

$$c_{ij}=\frac{\kappa\left(x_i,x_j\right)}{\left\|\kappa\left(x_i,x_i\right)\right\|\cdot\left\|\kappa\left(x_j,x_j\right)\right\|},i=1,2,\cdots,N;\ j=1,2,\cdots,N \tag{3}$$

由上述表达式可知，如果引用高斯核函数，则

$$c_{ij}=\kappa\left(x_i,x_j\right) \tag{4}$$

同样，如果引用多项式核函数或其他的核函数，则特征空间中每个样本向量的长度标准化为 1，也可以得到 $c_{ij}=\kappa\left(x_i,x_j\right)$。那么，核矩阵在某种程度上是特征空间中余弦矩阵的一种特殊形式，因此，理论上可以用余弦矩阵来学习核矩阵及其参数[8]。如果 $c_{ij}\to 0$，则特征空间中的向量很可能相互正交，它们的类间散度和类内散度都很大[12]，该情形下，很难正确分开样本数据。如果 $c_{ij}\to 1$，则特征空间中的所有样本向量将会集中于一点，或者位于同一直线上，显然该情形也不利于样本分类。

步骤 2：定义通用熵如下：

$$\mathrm{Entro}\left(c_{ij}\right)=-\sum_{i=1}^{N}\sum_{j=1}^{N}\left|c_{ij}\right|\log\left|c_{ij}\right| \tag{5}$$

将(4)式代入(5)式得到：

$$\mathrm{Entro}\left(c_{ij}\right)=-\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\kappa\left(x_i,x_j\right)\right|\log\left|\kappa\left(x_i,x_j\right)\right| \tag{6}$$

根据前面提到的两种情况 $c_{ij}\to 0$ 和 $c_{ij}\to 1$，通用熵 $\mathrm{Entro}\left(c_{ij}\right)$ 将会取到最小值 0，这将妨碍样本的分类。因此本文选取通用熵的最大值对应的参数建立核函数，便于对数据进行分类。这就是一个凸优化问题：

$$\mathrm{Max}\ \mathrm{Entro}\left(c_{ij}\right)$$

$$\text{s.t.}\begin{cases}0.1\le c\le 10.0\\ 1\le d\le 10\\ d\text{为整数}\end{cases}$$

$$\text{或}1.0\le\theta\le 10.0$$

4.3. 确定基本模式

有两种常见的确定基本模式的方法，一种是前向选择算法，另一种是后向选择算法[11]。考虑时间复杂度，本文选用前向选择算法，耗时较小。特征抽取的一项基本原则是，由不同投影方向得到的特征之间应尽量互不相关，因此，确定基本模式的要求是由两个基本模式得到的特征集之间具有较小的特征相关，这样特征抽取结果的贡献率比较大，综合信息的能力较强。利用核矩阵列向量 $v_i=\left(k\left(x_1,x_i\right),k\left(x_2,x_i\right),\cdots,k\left(x_N,x_i\right)\right)^{\mathrm{T}}$ 之间的特征相关即核矩阵列向量之间的余弦距离，来确定基本模式。两个核矩阵列向量 v_i 和 v_j 之间的余弦距离为

$$\cos\left(v_i,v_j\right)=\frac{\left(v_i,v_j\right)}{\left\|v_i\right\|\cdot\left\|v_j\right\|}$$

确定基本模式的过程如下：

步骤 1：确定前两个基本模式。

找出两个核矩阵的列向量 v_m 和 v_n 满足

$$\cos\left(v_m, v_n\right) = \min_{1\le i,j\le N} \cos\left(v_i, v_j\right)\left(1 \le m, n \le N\right) \tag{7}$$

然后，将它们对应的原始训练样本 x_m 和 x_n 作为前两个基本模式，重新标记为 x_1' 和 x_2'，将它放入新的集合 A 中，$A = \left\{x_1', x_2'\right\}$，构成新的矩阵 $K_2 = \begin{bmatrix} k\left(x_1, x_1'\right) & k\left(x_1, x_2'\right) \\ k\left(x_2, x_1'\right) & k\left(x_2, x_2'\right) \\ \vdots & \vdots \\ k\left(x_N, x_1'\right) & k\left(x_N, x_2'\right) \end{bmatrix}$。

步骤 t：确定第 t 个基本模式 $\left(3 \le t \le s\right)$。

假设已经在前面 $t-1$ 步中，确定了基本模式 $x_1', x_2', \cdots, x_{t-1}'$，并且 $A = \left\{x_1', x_2', \cdots, x_{t-1}'\right\}$，由这些基本模式构成了下面的新矩阵：

$$K_{t-1} = \begin{bmatrix} k\left(x_1, x_1'\right) & \cdots & k\left(x_1, x_{t-1}'\right) \\ k\left(x_2, x_1'\right) & \cdots & k\left(x_2, x_{t-1}'\right) \\ \vdots & \ddots & \vdots \\ k\left(x_N, x_1'\right) & \cdots & k\left(x_N, x_{t-1}'\right) \end{bmatrix}$$

对于任意 $x \in X - A$，设候选的核矩阵列向量是 v_p，

$$v_p = \left(k\left(x_1, x_p\right), k\left(x_2, x_p\right), \cdots, k\left(x_N, x_p\right)\right)^{\mathrm{T}}.$$

定义候选向量 v_p 与矩阵 K_{t-1} 之间的余弦距离为

$$\operatorname{cosdis}\left(v_p, K_{t-1}\right) = \frac{1}{t-1}\sum_{q=1}^{t-1} \cos\left(v_p, v_q\right) \tag{8}$$

式中，v_q 是 K_{t-1} 的第 q 列向量：$v_q = \left(k\left(x_1, x_q'\right), k\left(x_2, x_q'\right), \cdots, k\left(x_N, x_q'\right)\right)^{\mathrm{T}}$。利用式(7)确定 $\operatorname{cosdis}\left(v_p, K_{t-1}\right)$ 的最小值，与最小值对应的原始训练样本就确定为第 t 个基本模式，记为 x_t'，将它加入到集合A中。将样本 x_t' 对应的核矩阵列向量加到 K_{t-1} 作为该矩阵的最后一列，变为新矩阵 K_t。

重复上述过程直到 $t = s$，就确定了所有的基本模式 $x_1', x_2', \cdots, x_s'\left(s < N\right)$ 以及新的矩阵

$$K_s = \begin{bmatrix} k\left(x_1, x_1'\right) & \cdots & k\left(x_1, x_s'\right) \\ k\left(x_2, x_1'\right) & \cdots & k\left(x_2, x_s'\right) \\ \vdots & \ddots & \vdots \\ k\left(x_N, x_1'\right) & \cdots & k\left(x_N, x_s'\right) \end{bmatrix}_{N\times s}。$$

4.4. 重建 KPCA 模型

将训练样本 x_i 映射到高维之后，对映射后的数据 $\Phi\left(x_i\right)$ 进行中心化处理，即均值满足

$$\bar{\Phi} = \frac{1}{N}\sum_{i=1}^{N} \Phi\left(x_i\right) = 0$$

我们知道将一个向量 x_i 在 v 方向上的投影，若投影方向的模长 $\|v\| = 1$，则投影后的向量为 $\left\langle x_i, v\right\rangle v$。这就意味着在单位向量上投影后的坐标为 $\left\langle x_i, v\right\rangle = x_i^{\mathrm{T}} v = v^{\mathrm{T}} x_i$。那么特征抽取要求投影之后的数据尽可能互不相关，即要求找到一个投影方向 v，使得投影后的数据方差最大。投影后的方差为

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(v^{\mathrm{T}}x_i-\mu\right)^2=\frac{1}{N}\sum_{i=1}^{N}\left(v^{\mathrm{T}}x_i\right)^2=\frac{1}{N}\sum_{i=1}^{N}\left(v^{\mathrm{T}}x_i\right)\left(v^{\mathrm{T}}x_i\right)^{\mathrm{T}}\\ &=\frac{1}{N}\sum_{i=1}^{N}v^{\mathrm{T}}x_ix_i^{\mathrm{T}}v=v^{\mathrm{T}}\left(\frac{1}{N}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}\right)v=v^{\mathrm{T}}Cv\end{aligned} \tag{9}$$

上式中，假设所有样本已经经过中心化处理，$\mu=0$ 为投影点的均值，$C=\frac{1}{N}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}$ 为原训练样本的协方差矩阵则找出最优投影方向的问题可转化为下面的优化问题：

$$v=\arg\max v^{\mathrm{T}}Cv \tag{10}$$

$$\text{s.t. } \|v\|=1$$

为解决该问题，引入拉格朗日乘子 λ，建立拉格朗日函数

$$f(v,\lambda)=v^{\mathrm{T}}Cv-\lambda\left(v^{\mathrm{T}}v-1\right) \tag{11}$$

令 $f(v,\lambda)$ 对 v 和 λ 的偏导数为零得

$$Cv=\lambda v \tag{12}$$

$$v^{\mathrm{T}}v=1 \tag{13}$$

将(12)式和(13)式代入(10)式可转为

$$v=\arg\max\lambda \tag{14}$$

那么求解该优化问题可转化为求解协方差矩阵 C 的最大特征值。因此，对高维特征空间进行特征抽取时可转化为求解高维空间的协方差矩阵 D (D 为原始训练样本映射到高维空间后的协方差矩阵)的前 p (p 为主成分的个数)个最大特征值。高维特征空间的协方差为：

$$D=\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)\Phi(x_i)^{\mathrm{T}}=\frac{1}{N}\left[\Phi(x_1)\ \cdots\ \Phi(x_N)\right]\begin{bmatrix}\Phi(x_1)^{\mathrm{T}}\\ \vdots\\ \Phi(x_N)^{\mathrm{T}}\end{bmatrix}$$

对 D 进行特征分解的特征方程为

$$\lambda\boldsymbol{\alpha}=D\boldsymbol{\alpha} \tag{15}$$

式中，$\boldsymbol{\alpha}$ 为该特征方程非负特征值对应的特征向量。

由于直接对该协方差矩阵进行特征分解无法得到具体特征值，可以转化为对已确定的基本模式构建的核矩阵 K_2 进行特征分解得到特征值。其中，

$$\begin{aligned}K_2&=\begin{bmatrix}k(x_1,x_1')&\cdots&k(x_1,x_s')\\ k(x_2,x_1')&\cdots&k(x_2,x_s')\\ \vdots&\ddots&\vdots\\ k(x_s,x_1')&\cdots&k(x_s,x_s')\end{bmatrix}_{s\times s}=\begin{bmatrix}\Phi(x_1')^{\mathrm{T}}\Phi(x_1')&\cdots&\Phi(x_1')^{\mathrm{T}}\Phi(x_s')\\ \Phi(x_2')^{\mathrm{T}}\Phi(x_s')&\cdots&\Phi(x_1')^{\mathrm{T}}\Phi(x_s')\\ \vdots&\ddots&\vdots\\ \Phi(x_s')^{\mathrm{T}}\Phi(x_s')&\cdots&\Phi(x_1')^{\mathrm{T}}\Phi(x_s')\end{bmatrix}\\ &=\begin{bmatrix}\Phi(x_1')^{\mathrm{T}}\\ \vdots\\ \Phi(x_s')^{\mathrm{T}}\end{bmatrix}\left[\Phi(x_1')\ \cdots\ \Phi(x_s')\right]\end{aligned} \tag{16}$$

$$\text{令 } D=\frac{1}{N}\Phi(x_j)\Phi(x_j)^{\mathrm{T}}, \tag{17}$$

$$令\ K_1=\left(K_s\right)^{\mathrm{T}}=k\left(x_i',x_j\right)=\Phi\left(x_i'\right)^{\mathrm{T}}\Phi\left(x_j\right), \tag{18}$$

$$令\ K_2=\Phi\left(x_i'\right)^{\mathrm{T}}\Phi\left(x_r'\right), \tag{19}$$

其中，$i,r=1,2,\cdots,s,j=1,2,\cdots,N$。

将(17)式代入特征方程(15)可以得到：

$$\lambda\boldsymbol{\alpha}=\frac{1}{N}\Phi\left(x_j\right)\Phi\left(x_j\right)^{\mathrm{T}}\boldsymbol{\alpha} \tag{20}$$

将上述等式两边同时左乘$\Phi\left(x_i'\right)^{\mathrm{T}}$，右乘$\Phi\left(x_i'\right)$，得：

$$\lambda\left(\Phi\left(x_i'\right)^{\mathrm{T}}\Phi\left(x_r'\right)\right)\boldsymbol{\beta}=\frac{1}{N}\left(\Phi\left(x_i'\right)^{\mathrm{T}}\Phi\left(x_j\right)\right)\left(\Phi\left(x_j\right)^{\mathrm{T}}\Phi\left(x_i'\right)\right)\boldsymbol{\beta} \tag{21}$$

其中，$\beta=\left[\beta_1,\beta_2,\cdots,\beta_s\right]^{\mathrm{T}}$为新特征方程(21)式的特征向量。

将(18)式和(19)式代入(21)式，得：

$$\lambda K_2\boldsymbol{\beta}=\frac{1}{N}K_1\left(K_1\right)^{\mathrm{T}}\boldsymbol{\beta} \tag{22}$$

求解方程(22)得到前p个最大特征值$\lambda_t\left(t=1,2,\cdots,p\right)$以及其对应的特征向量$\boldsymbol{\beta}_t^*\left(t=1,2,\cdots,p\right)$。因此，原始训练样本的映射在高维空间特征向量上的投影(即主成分)可以如下表示：

$$B=\left[\frac{\sum_{i=1}^{s}\beta_{1i}^*k\left(x_i',x\right)}{\sqrt{\lambda_1}},\frac{\sum_{i=1}^{s}\beta_{2i}^*k\left(x_i',x\right)}{\sqrt{\lambda_2}},\cdots,\frac{\sum_{i=1}^{s}\beta_{pi}^*k\left(x_i',x\right)}{\sqrt{\lambda_p}}\right]^{\mathrm{T}} \tag{23}$$

其中，β_{1i}^*是向量$\boldsymbol{\beta}_t^*$的第i个分量[11]，s为已确定的基本模式的个数。

4.5. 多元线性回归预测

在高维特征空间中抽取了样本的p个主成分，对这p个主成分建立多元线性回归模型，如下：

$$y=\sum_{k=1}^{p}\omega_kB_k+\varepsilon \tag{24}$$

其中，y是股票收益率，ω_k为回归系数，B_k为主成分矩阵B的第k个分量，ε为残差矩阵。利用最小二乘法进行回归参数估计，得到

$$\hat{\omega}=\left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}}y \tag{25}$$

再将上式结果进行因变量y的预测为

$$\hat{y}=\sum_{k=1}^{p}\hat{\omega}_kB_k$$

利用方差分析法中的F统计量的P值(即显著性值)对回归方程进行显著性检验，一般P值小于0.05表示回归方程具有统计学意义。

5. 实证分析及实验结果

本文选用上证180成分股作为研究对象，选用前文所述的50多个因子进行基于EKPCA算法的多因子选股模型实证分析。其中，基本面数据选取上市公司2013至2018年这六年财务报表的季度数据，技术指标数据选取2013至2018年的周行情数据。这些数据时间跨度大，范围广，层面宽，获取数据样本充足，构建的投资组合考虑方面越周到，数据获取遵循可靠性，实效性，全面性，精确性，可操作性等

原则。数据来源为东方财富网站 Choice 金融终端，该数据库较权威，数据采集质量高，实验结果可信度高。当某只股票的指标数据空缺较多时，认为该样本不具备研究价值并且直接剔除，对于其它空缺值，本文通过 MATLAB2016a 软件进行数据预处理。由于因子之间的量纲不同，需要先将这 50 多个因子标准化之后，再通过 SPSS 22.0 统计软件进行多重共线性检验，下表 3 为输出的结果：

Table 3. Collinearity test
表 3. 共线性检验

影响因子	共线性统计	
	容许	VIF
流通 A 股	0.005	212.185
总股本	0.004	238.606
总市值	0.024	41.174
股东权益	0.016	62.108
市盈率	0.379	2.64
市净率	0.073	13.776
市销率	0.131	7.651
市现率	0.441	2.27
每股净资产	0.026	38.864
股息率	0.218	4.59
Alpha	0.039	25.404
Beta	0.09	11.063
Sharpe	0.131	7.608
Treynor	0.534	1.873
Jensen	0.079	12.662
可决系数R^2	0.235	4.258
EPS	0.006	156.609
每股营业收入	0.121	8.235
每股息税折旧摊销前利润	0.009	112.911
ROE	0.061	16.396
ROA	0.028	36.207
销售净利率	0.192	5.201
ROIC	0.026	37.753
销售毛利率	0.147	6.794
净资产/营业总收入	0.011	94.02
息税前利润/营业总收入	0.015	64.638
经营活动净收益/利润总额	0.439	2.279

Continued

营业利润/利润总额	0.537	1.863
资产负债率	0.051	19.617
权益乘数	0.088	11.361
流动比率	0.051	19.552
速动比率	0.037	26.939
现金比率	0.069	14.413
总资产周转率	0.125	7.991
营业周期	0.059	17.023
每股收益同比增长率	0.375	2.668
净利润同比增长率	0.139	7.17
营业利润同比增长率	0.136	7.374
现金流量净额同比增长率	0.426	2.348
营业收入同比增长率	0.205	4.872
净资产同比增长率	0.19	5.257
成交量	0.213	4.701
换手率	0.313	3.198
MACD 指数平滑异同平均	0.298	3.358
DMI 趋向指标	0.4	2.499
RSI 相对强弱指标	0.052	19.171
ROC 变动速率	0.157	6.367
KDJ 随机指标	0.103	9.668
OBV 能量潮	0.209	4.779
STD 标准差	0.036	27.471
预测 PEG	0.464	2.157
预测净利润增长率	0.373	2.682
预测营业收入增长率	0.375	2.667
PSY 情绪指标	0.207	4.82
BRAR 人气意愿指标	0.209	4.795

由表 3 可见，最大方差膨胀因子为 238.606 远远大于 10，认为这些因子之间的多重共线性比较严重，不能真实反映股票的收益率走向。因此，利用 EKPCA 算法模型进行降维处理，首先，找到最大通用熵对应的多项式核参数 c 和 d 分别是 1 和 4，高斯核参数 θ 为 8.2；然后，吴世农[15]等采用非回置式随机抽样方式确定上海股市适度的组合规模为 21~30 种股票，这一适度的组合规模可以较好地减少总风险，根据这一理论确定基本模式为 30 只股；最后，将这 30 只股票映射到高维空间进行特征抽取，对比两种核函数的特征抽取效率，发现多项式核函数只需抽取 4 个高效核主成分即可达到 95%的方差贡献率，而

高斯核函数需要抽取 21 个高效核主成分才能达到同样的效果，因此，本文采用多项式核函数建立核函数进行特征抽取。经过 MATLAB 2016a 程序输出结果如下：

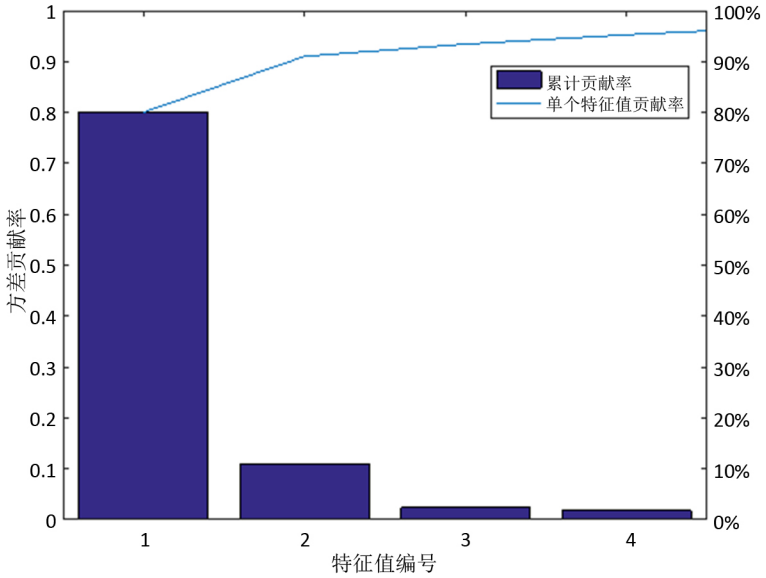

Figure 1. Variance contribution histogram and cumulative contribution histogram
图 1. 方差贡献率直方图与累计贡献率折线图

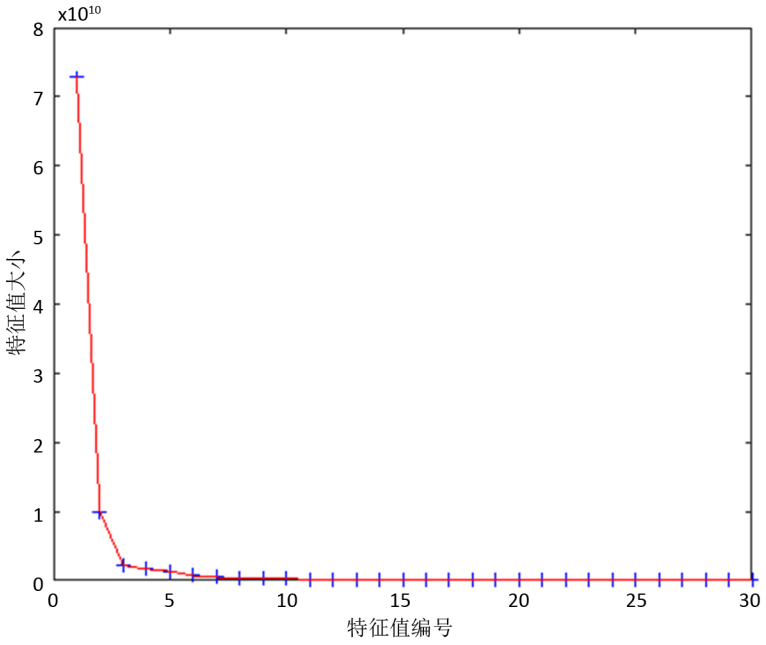

Figure 2. Eigenvalue lithotripsy
图 2. 特征值碎石图

由图 1 和图 2 可知选取前 4 个最大的特征值，它们的累计方差贡献率可以达到 95%，具有较强的综合因子差异的能力。再对比 EKPCA 算法与 KPCA 算法的特征抽取时间，发现 KPCA 算法完成特征抽取平均需要 0.1158 秒，而 EKPCA 算法平均只需 0.0154 秒，这表明 EKPCA 算法确实提高了特征抽取效率。

将确定的 30 个基本模式对应的收益率与抽取的 4 个高效核主成分进行以收益率为因变量，高效核主

成分为自变量的多元线性回归分析，预测基本模式的收益率。收益率的计算公式为

$$y = \frac{p_1 - p_0}{p_0}$$

其中，p_1为当期收盘价，p_0为前期收盘价，本文使用月收益率效果更好。通过 SPSS 22.0 输出回归结果如表 4 所示：

Table 4. Parameter estimation of multiple linear regression
表 4. 多元线性回归的参数估计

模型		非标准化系数		标准系数	t	显著性	共线性统计	
		B	标准错误	贝塔			容许	VIF
1	(常量)	6.722	2.038		3.298	0		
	高效核主成分 1	−1.54E−07	0	−0.083	−0.432	0.027	0.996	1.004
	高效核主成分 2	3.82E−06	0	0.19	0.99	0.003	0.995	1.005
	高效核主成分 3	3.19E−06	0	0.152	0.793	0.73	0.996	1.004
	高效核主成分 4	−3.95E−06	0	−0.127	−0.664	0.872	0.996	1.004

根据方差分析表中 F 检验对应的 P 值(显著性)为 0.012 < 0.05，说明至少有一个自变量能够有效预测因变量，回归模型具有实际应用价值。

再根据吴世农[14]等人的理论，本文选用收益率最高的前 25 只股票进行等权重分配的投资组合。投资组合结果如表 5 所示：

Table 5. Investment portfolio
表 5. 投资组合

排名	证券代码	证券名称
1	603019.SH	中科曙光
2	603799.SH	华友钴业
3	600025.SH	华能水电
4	600298.SH	安琪酵母
5	601360.SH	三六零
6	600519.SH	贵州茅台
7	600977.SH	中国电影
8	600352.SH	浙江龙盛
9	600900.SH	长江电力
10	601933.SH	永辉超市
11	600516.SH	方大炭素
12	601985.SH	中国核电
13	600061.SH	国投资本
14	601233.SH	XD 桐昆股
15	600585.SH	海螺水泥

Continued

16	601828.SH	美凯龙
17	600177.SH	雅戈尔
18	600688.SH	上海石化
19	600158.SH	中体产业
20	600518.SH	康美药业
21	600643.SH	爱建集团
22	600705.SH	中航资本
23	600011.SH	华能国际
24	600028.SH	中国石化
25	600598.SH	北大荒

6. 模型评价

好的投资组合具有高收益低风险的特征，因此，对已确定的组合投资进行绩效评价涉及对超额收益率以及风险的检测。本文将 2018 年底所有上证 180 的股票数据代入模型进行回测检验，收益率采用周行情数据，由 MATLAB 2016a 输出的图 3 及结果，发现上证 180 在该时间段中的平均收益率为 1.19%，而本模型确定的投资组合的平均收益率达到 1.75%，说明构造的该投资组合的收益情况较好。再用 β 系数和夏普比率对组合的风险进行评价，β 系数用来衡量投资组合相对于整个股市的价格波动情况，其值越大说明风险越大。夏普比率可以同时对收益以及风险加以综合的指标，其值越大，说明投资组合单位风险所获的风险回报越高。经过计算，投资组合的评价情况如表 6 所示。

由表 6 数据可以得出，该投资组合在所选时间段内系统风险小于大盘，且收益绩效大幅度超过大盘，可以获得较高的超额回报。

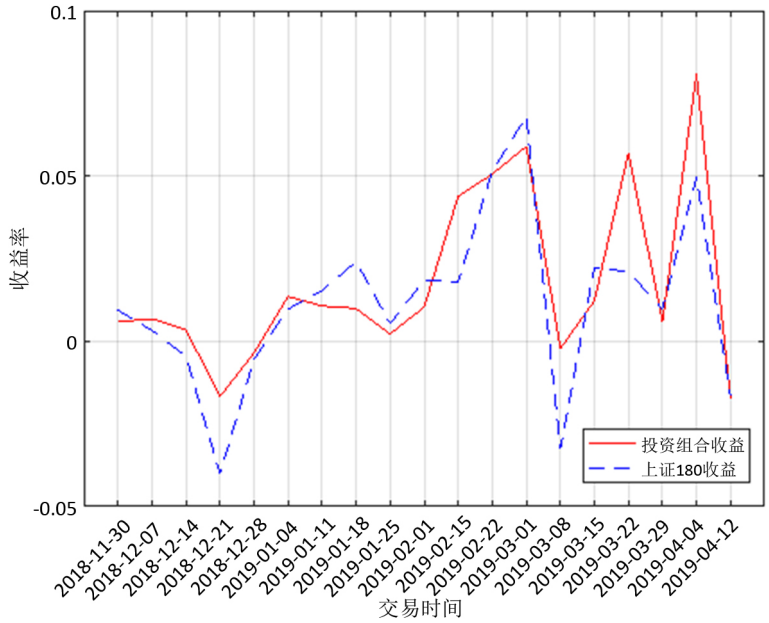

Figure 3. Portfolio performance analysis
图 3. 投资组合绩效分析

Table 6. Portfolio risk return situation
表 6. 投资组合风险收益情况

指标	市场基准水平	构建的投资组合
β系数	1	0.7787
夏普比率	1.302	2.1892

7. 总结

本文主要研究了基于 EKPCA 算法的多因子量化选股模型，该模型通过特征相关即核矩阵列向量之间的余弦距离来确定基本模式，不仅提高了以往学者的 KPCA 算法的特征抽取效率，而且减少了存储空间。从识别角度来看，EKPCA 的识别效果也是很好的，可以在训练样本规模较大时，高效地找出影响股票收益率显著因子，构造较好的投资组合。本文在研究多因子相关关系时，通过方差膨胀因子检验法检测发现所选多因子间存在严重的多重共线性问题，提出 EKPCA 算法进行降维。舍弃最小特征值对应的特征向量，仍然能够保持原样本低维结构的完整性，不仅使样本的采样密度增大，而且起到降噪的效果。

基金项目

江苏大学 2019 年大学生实践创新训练计划项目(项目编号：5553170019)。

参考文献

[1] 王春丽, 刘光, 王齐. 多因子量化选股模型与择时策略[J]. 东北财经大学学报, 2018, 119(5): 83-89.

[2] 范烨. 多因子选股模型建立的研究[J]. 全国流通经济, 2018(3): 64-65.

[3] 李娜, 毛国君, 邓康立. 基于 k-means 聚类的股票 KDJ 类指标综合分析方法[J]. 计算机与现代化, 2018, 278(10): 12-17.

[4] 苏治, 傅晓媛. 核主成分遗传算法与 SVR 选股模型改进[J]. 统计研究, 2013, 30(5): 54-62.

[5] 贾秀娟. 基于随机森林的支持向量机量化选股[J]. 区域金融研究, 2019(1): 27-30.

[6] 吕凯晨, 闫宏飞, 陈翀. 基于沪深 300 成分股的量化投资策略研究[J]. 广西师范大学学报(自然科学版), 2019, 37(1): 1-12.

[7] 徐景昭. 基于多因子模型的量化选股分析[J]. 金融理论探索, 2017(3): 30-38.

[8] 朱晨曦. 我国 A 股市场多因子量化选股模型实证分析[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2017.

[9] 凌士勤, 苏乐. 投资者情绪与股票收益的实证研究——基于扩展卡尔曼滤波的方法[J]. 时代金融, 2017(6): 192.

[10] 王锐. 岭回归分析在解决经济数据共线性问题中的应用[J]. 经济研究导刊, 2018(22): 144-147.

[11] Fan, Z., Wang, J., Xu, B. and Tang, P. (2014) An Efficient KPCA Algorithm Based on Feature Correlation Evaluation. *Neural Computing and Applications*, **24**, 1795-1806. https://doi.org/10.1007/s00521-013-1424-9

[12] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 60-62, 128.

[13] Sun, R., Tsung, F. and Qu, L. (2007) Evolving Kernel Principal Component Analysis for Fault Diagnosis. *Computers & Industrial Engineering*, **53**, 361-371. https://doi.org/10.1016/j.cie.2007.06.029

[14] 范自柱. 新型特征抽取算法研究[M]. 合肥: 中国科学技术大学出版社, 2016: 95-102, 122-128.

[15] 吴世农, 韦绍永. 上海股市投资组合规模和风险关系的实证研究[J]. 经济研究, 1998(4): 22-29.

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-0967，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：fin@hanspub.org