基于主成分分析与因子分析的开放式公募基金评价模型

潘佳祥　董傲坤　鲍一丹

（天津科技大学，天津 300450）

摘要：随着公募基金数量与基金管理人的增加，基金市场也存在基金业绩分化的现象。为此，本文基于主成分分析与因子分析建立基金业绩评价与推荐模型，从基金的收益水平和风险程度两方面进行综合分析。首先，通过计算基金的日均收益率、超额收益率和夏普比率作为基金的收益指标，通过计算基金的最大回撤值、标准差作为基金的风险指标。其次，使用 SPSS 软件，对指标进行主成分分析、因子分析，构建因子得分矩阵计算综合得分排名，并作出推荐。最后，通过 matlab 工具箱建立建立多元线性拟合模型对基金未来一年的表现进行预测。

关键词：主成分分析；因子分析；基金评价；SPSS；Python

中图分类号：F832.51　　**文献识别码**：A　　**文章编号**：2096-3157（2021）31-0142-03

DOI:10.16834/j.cnki.issn1009-5292.2021.31.045

基金投资在近些年来是一个热门话题，越来越多的投资者将购买基金作为投资方式，截至 2020 年第四季度末，我国的基金数量已多达 7000 多只。同时我国的开放式公募基金经过了 20 多年的发展，截至 2020 年年底，其规模已经跃居亚太地区第一，全球第五。并且，随着开放式公募基金的发展，出现了越来越多的基金经理，其管理的基金规模也在逐步扩大。但是，由于基金经理的水平差异较大，基金之间的业绩也各不相同，基金投资领域出现分化的现象。许多投资者面对如此众多的开放式公募基金与基金经理，很难作出最优选择。

面对数量众多的基金与投资者难以选择的局面，开放式公募基金的业绩评价和推荐模型对基金销售机构和广大投资者都具有重要意义。

一、模型的建立与分析

1. 数据来源

本文利用基于 Python 的开源金融数据接口库收集 2008 年至 2020 年所有开放式公募基金从成立以来的所有历史净值数据，包括每日净值、累计净值和收益率，并进行数据清洗，得到有效数据。

2. 主成分分析与因子分析概述

主成分分析与因子分析的主要思想类似，二者都是利用了“降维”思想，利用若干主成分或者公因子来代表原始变量。

主成分分析是利用线性的变换方法，把高维的问题向低维的问题转化，把研究变量通过降维处理，转变成新的变量，同时按照方差降序排列。主成分分析的方法可以将很多变量通过降维处理，转化成少数互不相关的新变量。所形成的新变量能够解释原有数据的绝大多数变量，即原有变量的主要成分，并且能够解释数据的综合指标。因子分析同样运用了“降维”理念，结合原有变量的相关矩阵内部关系，将关系复杂变量转化为少数公共因子和特殊因子线性组合。相比较于主成分分析法，因子分析则侧重于描述原始变量之间的相关关系。

3. 模型的建立

（1）收益率指标

基金的收益水平是基金的业绩的重要组成部分。本文选取日均收益率、超额收益率两个收益评价指标来对基金收益情况进行衡量。首先建立基金业绩与收益水平的关系函数。设第 i 只基金的收益率为 r_i，第 i 只基金的单位净值为 p_i，基金份额为 X，P_0，P_1 分别为 0 时刻与 1 时刻基金单位净值，

日均收益率：$\overline{r_n}=\dfrac{X(1-f)\sum_{i=1}^{n}P_n-\sum_{i=1}^{n-1}P_{n-1}}{n}$

超额收益率：$R_{it}=r_i-r_m$

（2）风险衡量指标

投资中的风险问题也不容忽视，主要是基金投资收益的不确定性，基金经理一般会结合自己独有的投资经验与操作风格，通过各种方式来达到规避风险的目的，可是基金的收益波动不能被完全消除。本文选用基金收益率的标准差、最大回撤值这两个指标衡量基金的风险情况。

标准差：$\sigma=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(r_i-r_m)^2}$

最大回撤值：$drawdown=Max\left(\dfrac{P_n-P_{n+1}}{P_n}\right)$

（3）夏普比率

基金投资中，风险的差异往往意味着收益的差异。夏普比率同时将风险因素与收益因素综合考虑，能够减少因考虑单一风险或收益因素来评价投资的误差，提高评价的准确性与客观性。因此，本文选取夏普比率作为综合评价指标。

夏普比率：$S=\dfrac{\overline{r_n}-r_f}{\sigma}$

由于我国国债市场流动性较弱，利率市场化较弱，不适合直接选取国债利率。因此本文综合考虑后，决定选用中国银行 2021 年度的一年期定期存款利率 1.75% 为投资无风险利率，即 r_f=1.75%。

（4）综合评价

最后运用 SPSS 软件对得到的相关数据进行主成分分析找出两个主成分，通过因子分析得到旋转后的方差贡献率作为因子权重。将因子权重与对应成分构建线性方程，得出综合得分并排序，从中选择前 10 的基金作为推荐基金。然后把这 10 只基金 2015 年每日收益率赋予相等的权重并求和，作为组合的收益

率，在此基础上对所推荐基金未来一年，即 2016 年的日收益进行预测。

4. 模型的分析

本文对 2008 年至 2020 年中每一年已成立的基金分别进行分析，在该年尚未成立的基金不予考虑。接下来从基金的收益水平和风险程度两方面进行综合分析。一方面，通过计算基金的日均收益率、超额收益率和夏普比率判断基金的收益水平；另一方面，通过计算基金的最大回撤值、标准差作为基金的风险指标。综合从以上五个角度进行分析。本文选取 2016 年基金的发展状况进行展示分析。

（1）分析一：相关性分析

对每年各只基金的日均收益率、超额收益率、夏普比率、最大回撤值、标准差这些指标进行相关性分析。使用 SPSS，对所选取指标进行相关性分析，绘制相关性热力图表示不同指标之间的相关程度，如图 1 所示。

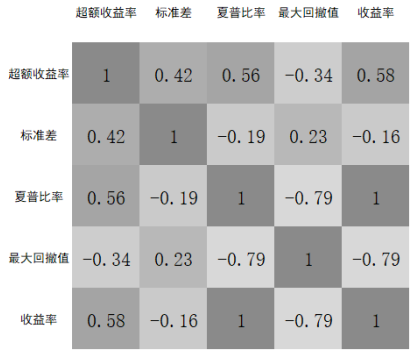

图1　相关系数矩阵热力图

注：一般情况下，相关系数r的绝对值在0.8以上，认为A和B有强的相关性；在0.3～0.8之间，可以认为有弱的相关性；在0.3以下，认为没有相关性。

从图 1 可得，由于收益率和夏普比率的相关系数为 1，说明这两个指标完全正相关。超额收益率与标准差、夏普比率、收益率有着较弱的正相关性，与最大回撤值有着较低的负相关性；夏普指数和收益率与最大回撤值有着较强的负相关性。因此，收益率和夏普比率可以作为一个指标，其他指标均具有一定的可解释性，可作为衡量基金收益或风险的指标。

（2）分析二：日均收益率

日均收益率，代表着基金成立以来的收益状况的一般水平，以基金代码为 160918 为例计算日均收益率约为 0.3657。同理可得该年其他基金日均收益率。通过 Python 编程筛选出 2015 年初已成立的基金，在 2015 年每天的收益率数据，计算各基金的日均收益率。

（3）分析三：超额收益率

超额收益率 $R_{it}=r_i-r_m$，反映一只基金的收益情况高出整个市场平均水平的程度。超额收益率越高就说明基金的收益情况越好。运用 Python 编程筛选出 2015 年年初已成立的基金，在 2015 年每天的超额收益率数据，进行求解。以 160918 的超额收益率计算为例：即 $R_{it}=r_i-r_m=0.3675-0.1216=0.2459$，同理得出各基金 2015 年 1 整年的超额收益率

（4）分析四：标准差

基金收益率的标准差代表基金收益的波动情况，标准差越大意味着基金收益情况越不稳定。运用 Python 编程筛选出 2015 年已成立基金在 2015 年 1 整年每天的标准差数据，得出各基金 2015 年 1 整年的标准差，以编号为 166016 的基金收益标准差为例进行计算。结果如下：

$$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i-r_m)^2}=\sqrt{\frac{1}{245}\sum_{i=1}^{n}(r_i-0.1216)^2}\approx 247.1761$$

（5）分析五：夏普比率

夏普比率表示投资者每多承担一单位风险，能够获得超过无风险报酬率的报酬。在夏普比率大于零的情况下，说明在衡量期内基金的平均净值增长率超过了无风险利率，夏普比率越大，表示投资该基金承担单位风险能够获得的收益越高。以代码为 160918 的基金夏普比率为例计算。结果如下：

$$S=\frac{\overline{r_n}-r_f}{\sigma}=\frac{0.3657-0.0175}{280.5581}\approx 0.0012$$

同理可得其他基金夏普比率。以根据 Python 编程所筛选出 2015 年初之前成立的所有基金的 2015 年 1 整年夏普比率。

（6）分析六：最大回撤值

最大回撤值是衡量在投资某一产品之后，可能会出现最大程度亏损的风险指标。最大回撤值越小说明基金的风险越低。以基金代码为 160918 的基金为例计算最大回撤值：即将该基金 2016 年每一天的当天单位净值与后一天的单位净值作差，再除以当天单位净值求出回撤值，然后求出一年中最大回撤值约为 -0.0036，同理可得其他基金最大回撤值。

二、模型的求解

1. 求解一：推荐合适的基金

为保证模型准确性，需保证使用至少一整年的数据进行计算分析，以 2016 年推荐基金为例，那么只需考虑截至 2015 年 1 月 1 日已成立的 223 只基金，在此范围中进行基金推荐。

考虑评价模型的客观性、有效性，本文使用数理统计中主成分分析法、因子分析法结合 SPSS 软件，分别对两类（风险与收益）的五个评价指标进行了综合的测评与排序。因为主成分 1、2 特征值都大于 1，而且能够解释原数据 89.42% 的方差，即涵盖了大部分信息，这表明前两个主成分能够代表最初的 5 个指标，所以本文选取前两个主成分的方式具有一定可取性。

表 1 和表 2 分别给出了 2016 年基金主成分、特征值、方差贡百分比与累计献率和经过旋转之后的成分矩阵。

表1　总方差解释

成分	特征值	方差百分比	累计贡献率%
1	3.072	61.449	61.449
2	1.399	27.97	89.420

表2　旋转后的成分矩阵a

指标	成分1	成分2
R_{it}	0.564	0.736
标准差	-0.241	0.914
夏普比率	0.98	0.048
最大回撤值	-0.881	0.116
收益率	0.979	0.07

旋转方法：凯撒正态化最大方差法；a 旋转在 3 次迭代后已收敛。

从旋转之后的成分矩阵可以看出，收益率、夏普比率和超额收益率和主成分 1 具有较大相关性，将其归纳整理，作为评价基金收益情况的相关指标，主要表示开放式公募基金的收益水平和业绩情况。最大回撤值和标准差与主成分 2 具有较大相关性，可以将其归纳总结为开放式公募基金的风险调控的相关能力指标。分别用 Y_1 和 Y_2 代表主成分 1 和主成分 2。

利用 SPSS 软件，将所求得的因子载荷矩阵中的数值与对其应特征根的算术平方根作商，求得主成分的线性组合系数，将

所得系数代入模型，求解出主成分的表达式。然后利用成分矩阵旋转之前的方差贡献率求得各自的权重，再将每一项的权重与所对应的主成分值相乘，计算结果作为综合得分。即：

Score=0.61449×Y_1+0.2797×Y_2（Y_1，Y_2 ∈ R）

最终得出 2016 年初推荐的十只基金依次为：大成中小盘混合 A、国泰估值优势混合、国泰中小盘成长混合、兴全轻资产混合、银华内需精选混合、申万菱信量化小盘股票、兴全合润混合、东方红睿丰混合、融通领先成长混合、万家行业优选混合 A。以排名第一的基金大成中小盘混合 A 综合得分为例，Y_1 为 2.8660，Y_2 为 0.4569，得分为 1.89。

将主成分综合评价结果与单一的收益水平和风险程度的评价结果进行 Pearson 相关性分析，结果显示可以通过相关性检验，说明主成分综合评价与所选取的单一指标具有较强相关性，因此能够使用主成分综合评价法进行基金的业绩评价。

2. 求解二：预测等权重组合基金未来一年的表现

在得出 2016 年年初推荐的 10 只优秀的基金后，还需要给出这 10 只基金等权重组合在 2016 一整年的的表现情况。所以将 2015 年中排名前 10 的每只基金的日收益率分别乘以 10% 的权重，再将其加总作为当日的组合收益率，最终得出基金等权重组合 2015 年整年间的每日组合收益率。

此外，由于基金市场节假日休市，所以一年中每只基金有效交易天数为 244 天。因此可以将 2015 年的有效交易日期用序号 1-244 表示，2016 年的有效交易日期则可用 245-494 表示。这样就把 2015 年和 2016 年的日期放于同一坐标轴，以 2015 年基金的每日组合收益为基准，运用 matlab 工具箱反三角函数拟合方法,得到函数 f(x),对 2016 年的日收益进行预测。函数 f(x) 的系数和变量的拟合程度分别如表 3 和图 2 所示。

$$f(x)=\sum_{i=1}^{7} a_n \sin(b_n x+c_n),(1 \leqslant x \leqslant 454 且 x \in N^*)$$

表4　模型系数

系数	数值	系数	数值	系数	数值
a_1	1.028	b_1	63.71	c_1	-2.08
a_2	0.8288	b_2	108.3	c_2	0.7377
a_3	0.7087	b_3	9.57	c_3	-0.5863
a_4	0.8604	b_4	111	c_4	0.5118
a_5	0.289	b_5	121.3	c_5	2.524
a_6	0.8634	b_6	66.14	c_6	-2.337
a_7	0.6745	b_7	22.54	c_7	-0.3295

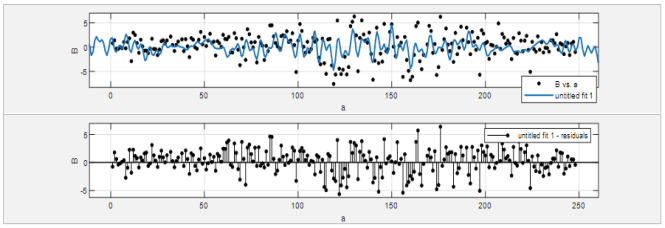

图2　模型拟合程度

通过计算结果以及拟合模型图所示，函数 f（x）具有较高的拟合程度，说明模型较为准确。因此，本文将运用该模型预测每年年初所推荐的十只开放式公募基金等权重基金组合在未来一整年的发展趋势。

三、结语

关于基金的评价，目前有很多学者从不同的角度建立模型，并进行分析求解，可是截止目前仍没有全面准确的基金评价模型，可见这一问题十分复杂。

本文对开放式公募基金评价这一问题进行探索，建立基于主成分分析与因子分析的开放式公募基金的评价模型。首先查阅大量参考文献，对影响基金发展的因素进行收集与分析后，选用基金的日均收益率、超额收益率和夏普比率、最大回撤值和标准差等 5 个重要指标。通过 SPSS 软件做相关性分析得到这些指标间具有一定的相关性后，基于各个基金的单位净值和日增长率建立 5 个指标的函数关系式。

以 2016 年为例，通过 python 编程分别筛选出 2015 年年初之前成立的所有基金 2015 年 1 整年内 5 个指标的相关数据。然后进行主成分分析、因子分析，得出旋转后的成分矩阵，将日均收益率、夏普比率和超额收益率划入收益型指标，将标准差和最大回撤值划入风险型指标。利用 SPSS 软件得到主成分的线性组合系数，由此得到主成分的表达式。再求得权重，再将每一项的权重与所对应的主成分值相乘，计算结果作为综合得分。取前 10 名作为 2016 年年初的推荐基金。在假设无风险收益率是定值、忽略现金分红的情况下，分别求解目标函数，得出各基金 2015 年 1 整年的 5 个指标的值,并根据 5 个指标分别排名，作为 2016 年年初基金的前十名。

最后将所推荐的十只基金等权重组合，对未来一年，即 2016 年的日收益进行预测。同时，模型还可以进行调整与推广，例如可以将开放式公募基金按照风险成分或收益成分的得分进行排名，根据投资者对于风险的承受能力与偏好程度，细分为保守型、激进型等多种类型投资者，为不同类型的投资者推荐合适基金。

所建立的基于主成分分析与因子分析的开放式公募基金评价模型能够为基金投资者提供一定参考，能够帮助基金投资者在众多基金中综合分析，选择最优基金。

参考文献：

[1] 何菊香，王徵羽．中国大数据基金的业绩评价与对比分析[J]．科技促进发展，2020，16（10）：1164-1173.

[2] 董艳青．我国开放式基金业绩评价的实证研究[D]．长沙：湖南大学，2008.

[3] 王元璋．基于因子分析法的区域物流能力评价研究：以河北省为例[J]．通化师范学院学报，2021，42（9）：96-100.

[4] 李杰，段小明，王艳萍．因子分析在上市公司财务绩效评价中的应用[J]．吉林化工学院学报，2021，38（7）：101-106.

[5] 付灿灿．股票型基金绩效评价研究[D]．上海：上海外国语大学，2021.

作者简介：

1. 潘佳祥，天津科技大学在读本科生；研究方向：金融工程。

2. 董傲坤，天津科技大学在读本科生;研究方向:人力资源管理。

3. 鲍一丹，天津科技大学在读本科生；研究方向：金融工程。