2021年第二届“大湾区杯”粤港澳金融数学建模竞赛

题 目 基于多因子模型的量化选股策略

摘 要：

本文主要聚焦于多因子的选股策略，建立多因子选股模型，首先提取研报特征指标，构建因子库，并通过分配权重合并因子，最终得到给出相应投资策略，且要根据突发事件进行动态调整投资策略。

针对问题一，本文首先进行数据提取和预处理环节。在本环节中，我们将选用 ichoice 平台进行**研报基本数据**的采集，并对其进行中位数去极值、数据标准化、缺失值处理等初步的数据处理，最终构建出完善的备选因子数据库。

针对问题二，可分为四个部分：因子有效性检验、因子相关性检验、相关小因子合并和多因子模型构建。对于因子有效性检验采用单因子 IC 分析法来判断因子与收益率的相关强度，并确定了所选因子的合理性。对于那些高度相关的因子，本文并非简单将同类型中表现较差的因子剔除，而是采取**主成分分析法**，将相关系数较大的小类因子合并为一个大类因子，从而扩大数据的宽度，使因子的综合收益贡献率提高。对于大类因子采用 **IC 加权**的方法构建最终选股模型。最终得到的投资策略回测时年化收益率达 32.58%。

针对问题三，针对在真实的股票市场之中，很多具有影响，但没有被考虑进传统选股模型之中的因素，例如以突发事件闪现、舆情影响、自然灾害影响为主的外部环境因素，通过查阅相关文献，确定这些事件中的**不可预测性**事件的**正向与负向**影响，对在问题二环节中建立的模型中各个因子的影响，从而调整各个因子在模型中的权重。

针对问题四，将问题三中的因子权重调整加以实现，得到考虑外部因子后的优化模型。后在观测期对模型进行回测，得到模型最终的组合收益率，并确定投资选股策略。

关键词：研报数据；多因子选股；主成分分析；IC 加权；不可预测事件

目录

1 问题重述

根据有效市场理论，市场在处于强有效状态时，股票包含了所有的市场信息，投资者无法从以往的价格中获得更多信息从而达到套利的目的。但现实是，市场远没有理论上的透明，一般认为，中国股市处于半强式有效市场和弱式有效市场之间。也就是说，仍有从那些未反映到股价中的信息中获取超额利润的空间。

而券商研报（卖方研报）正是沟通企业与市场的桥梁，让企业信息快速向市场传递，使企业的经营状况更快的反映到股价中。券商研报提供的丰富且及时的信息，是投资进行决策的重要参考资料。但面对研报包含的海量信息，如何建立起它们与收益率之间的关系，从而跑赢市场，获得超额收益率是量化投资领域的核心难题。

而在量化投资策略中应用最广泛的模型之一是多因子模型，其基本原理是利用一系列的有效因子构建模型，并利用构建好的模型预测股票下期收益率从而选择出表现优越的股票。多因子主要有三种形式：宏观经济因子模型、基本面因子模型、统计因子模型。其中，基本

面因子模型使用可观察到的股票自身的基本属性作为解释股票市场收益率变动的变量，其效果远好与其他两类模型，是业界主流的研究方向，也是本题题干中要求我们探索的主体。

基于上可知，题目要求我们构建相关模型解决以下问题：

（1）提取研报中特征指标，构造出各类特征因子，构建备选因子库。

（2）对提取出的特征因子进行有效性检验和相关性检验，筛选出有效因子并进行因子合并。最后计算出各因子权重，构建多因子模型，给出投资策略。

（3）考虑外部环境影响（突发事件闪现、舆情影响、自然灾害影响），探究其对原有模型中各因子的影响

（4）在（2）搭建的多因子模型上将（3）中形成的外部环境因子加入考虑，对模型进行改进，给出最终的投资策略。

2 问题分析

本文要解决的是多因子模型的构建与改进问题。问题一要求我们提取研报特征指标，构建备选因子库；问题二要求我们选定有效因子，构建多因子模型，给出初步投资策略；问题三要求我们考虑外部环节对模型内因子的权重影响；问题四要求我们综合考虑问题二三构建的模型，给出相应投资策略。

2.1 问题一的分析

本文认为，问题一是构建多因子模型的前置预处理环节。在本环节中，我们将选用 ichoice 平台进行研报基本数据的采集。并对其进行中位数去极值、数据标准化、缺失值处理等初步的数据处理，最终构建出完善的备选因子数据库。

2.2 问题二的分析

问题二可分为四部分：因子有效性检验、因子相关性检验、相关小因子合并和多因子模型构建。

首先，对于因子有效性检验，本文认为可以采用单因子 IC 分析法来判断因子与收益率的相关强度。IC 即信息系数，表示所选股因子值与股票下期收益率在横截面上的相关系数。通过 IC 值可以判断因子值对下期收益率的预测能力，IC 的绝对值越大，该因子越有效。本

文对所选股的股票因子值与股票下期的实际回报率在横截面上建立回归，分别计算出因子的平均因子收益、IC 的均值以及 IR（信息比率），从而完成初步的有效性检验。

其次，因为相同类型的因子反映的经济规律往往是一致的。因此，我们将个股因子值构建相关矩阵，对初步入选的因子进行相关性检验，得到因子之间的相关系数。

值得注意的是，对于那些高度相关的因子，本文并非简单将同类型中表现较差的因子剔除，而是采取主成分分析法，将相关系数较大的小类因子合并为一个大类因子，从而扩大数据的宽度，使因子的综合收益贡献率提高。

最后，在完成了因子的合并之后，就应着手搭建多因子模型。本文将采取主成分分析法方法分析各因子 IC 值从而确定出权重，给出投资方案。

2.3 问题三的分析

为了模型的准确与稳定，在构建多因子选股模型的时候往往只能考虑到本身具有时序特征的数据，以它们为基本构建因子。但在真实的股票市场之中，还有很多影响因素没有被考虑进模型之中。例如以突发事件闪现、舆情影响、自然灾害影响为主的外部环境因素，当它们发生时，会对股票行情产生较大的影响，因此，应该将这些因素纳入模型的考虑范围。

本文通过查阅相关文献，确定突发事件闪现、舆情影响、自然灾害影响对在问题二环节中建立的模型中各个因子的影响，从而调整各个因子在模型中的权重，以达到将外部环境因素纳入考虑的目的。

2.4 问题四的分析

在第四问，将问题三中的因子权重调整加以实现，得到考虑外部因子后的优化模型。后在观测期对模型进行回测，得到模型最终的组合收益率，并确定投资选股策略。

3 基本假设

1.中国股票市场为非强有效市场；

2.券商研报非预测部分数据准确无误；

3.所研究的事件之间没有重叠或冲突。

4 符号说明

符号	说明
X_{mean}	X 因子的平均值
X_{mean}±nσ	X 因子与平均值的距离
μ	样本数据的均值
σ	样本数据的标准差
IC^T	因子在 T 期的 IC 值
r^{T+1}	第 T+1 期的收益率
X^T	第 T 期因子 X 的暴露度
ω_{IC}^i	第 i 个细分因子在大类因子中分配的权重
$\boldsymbol{\rho_i}$	第 i 个细分因子的贡献率
IC_{mean}^t	第 t 大类因子的平均 IC
IC_{mean}^i	第 t 大类下 n 个细分因子的 IC 均值
ω_{IC}^t	第 t 大类因子的权重

5 多因子选股模型的构建与检验

5.1 问题一：数据预处理与备选因子库构建

在多因子选股模型之中，最为重要的步骤之一就是选定有效的、稳定的因子。本文在参考多因子模型的研究后，根据市场经验与投资逻辑，认为选取候选因子应符合以下三个标准：

（1）普适性。不同上市公司处于不同的行业，必然存在一些因子不具有现实意义或不能够计算，随意候选因子应能够在不同行业、不同市值的上市公司中均适用。

（2）相关性。候选因子应能够比较好的显示出上市公司股票在考察期中的收益，根据相关因子制定的投资策略应大概率跑赢市场，获得稳定的超额收益。

（3）稳定性。候选因子对股票收益的影响应在长时段内持续展现，仅在短时间内有影响作用的因子不应被选为候选因子。

根据上述三个标准，本文选取了规模、价值、质量共三类因子。并在这三大类因子下选取细分因子：

规模类因子：指股票规模相关的因子。一般来说，市场认为规模较大的上市公司内部管理规范、流动性强且存在规模效应，持有这些公司的股票风险较低且在长期内能获得稳定的收益。而追求高收益率的投资者往往偏爱小市值的企业，认为其成长空间较大。通常选取总市值，流通市值作为细分因子。

价值类因子：指与股票估值相关的因子。由于市场存在局部无效性，有效信息无法完全的被反映在股价中，因此股票价格和内在价值可能存在较大的偏离。于是，投资者往往根据上市公司估值对公司发展前景进行合理预期。通常选择市净率（PB）、市销率（PS）、市盈率（PE）作为细分因子。

质量类因子：指与股票的财务质量、资本结构相关的因子。质量类因子能够反映公司发展的质量，展现公司发展的可持续性，是一类解释力较强的因子，往往受到投资者的重点关注，通常选择权益回报率（ROE）和资产回报率（ROA）为细分因子。

5.1.1 数据预处理：3σ 法去极值

3σ 法又称为标准差法。标准差本身可以体现因子的离散程度，是基于因子的平均值 X_{mean}而定的。在离群值处理过程中，可通过用 X_{mean}±nσ 来衡量因子与平均值的距离。标准差法处理的逻辑:

第一步：计算出因子的平均值与标准差

第二步：确认参数 n（这里选定 n = 3）

第三步：确认因子值的合理范围为 [X_{mean}−nσ, X_{mean} nσ]，并对因子值作如下的调整：

$$X_i' = \begin{cases} X_{mean} + n\sigma & \text{if } X_i > X_{mean} + n\sigma \\ X_{mean} - n\sigma & \text{if } X_i < X_{mean} - n\sigma \\ X_i & \text{if } X_{mean} - n\sigma < X_i < X_{mean} + n\sigma \end{cases}$$

其中， X_{mean}表示因子的平均值， X_{mean}±nσ 表示因子与平均值的距离。

5.1.2 数据预处理：缺失值处理

缺失值处理:本文采用股票多因子的截面数据，由于股票多因子是衍生计算而来，因此股票因停牌等原因导致的部分值缺失可能会影响到多个因子数据的完整新。同时本文用的股票包含众多行业，加大了存在某些指标缺失的可能性。大量缺失值的存在必然会影响模型的预测结果，甚至某些算法遇到数据缺失情况会直接报错导致训练过程都无法进行。因此在得到因子数据时，需要检查是否存在缺失值的情况。如果存在少量的缺失值则进行相应的缺失值填充处理:如果单支股票缺失值过多，则认为其不符合目前的选股条件，将该条股票因子数据剔除样本。为了保证模型训练有充足的样本数据，本文对于少量缺失值的股票样本并没有进行剔除，而是采用均值填充法对其进行填充。

5.1.3 数据预处理：标准化处理

标准化处理:在多因子模型中，由于各因子的性质不同，通常具有不同的量纲和数量级。当因子间的水平相差很大时，数值较大因子的作用就会被放大，而数值较小因子的作用就会被忽略掉，进而导致模型失效。因此为了保证模型的有效性，需要保证因子对模型的贡献程度相同，需要将因子值调整为具备相同的量纲和数量级的数据。

本文采用标准差标准化，是使用因子的均值和标准差对所有因子进行的标准化，进面得到一个近似服从 N(0.1)正态分布的新序列。计算公式如下所示，

$$X = \frac{x - \mu}{\sigma}$$

其中 μ 代表样本数据的均值，σ代表样本数据的标准差

5.1.4 处理后部分数据展示

通过上述步骤对原始数据的处理后，我们得到可用于构建因子的特征数据，以下是处理后的部分数据展示：

总市值	流通市	市净率	市销率	市盈率	权益回报率	资产回报率

	值	PB	PS	PE	ROE	ROA
-0.53641	-0.51869	-0.59563	-0.09853	-0.06983	-1.35828	-0.52353
-0.53604	-0.51832	-0.57668	-0.0724	-0.05194	-1.35828	-0.52353
-0.53613	-0.51841	-0.58125	-0.07871	-0.05626	-1.35828	-0.52353
-0.5361	-0.51838	-0.57995	-0.0769	-0.05502	-1.35828	-0.52353

5.1.5 备选因子数据库

大类因子	细分因子	英文缩写
规模类因子	总市值	Mkv
	流通市值	Neg_mkv
价值类因子	市净率	PB
	市销率	PS
	市盈率	PE
质量类因子	权益回报率	ROE
	资产回报率	ROA

5.2 问题二：因子检验、因子合并和因子权重计算

5.2.1 因子有效性检验：单因子 IC 分析法

IC 即信息系数，表示所选股票因子与下期收益率的界面相关系数。其取值范围为[-1,1]，当 IC 为正数时，表示所选股票因子与下期收益率正相关；当 IC 为负数时，表示所选因子与下期收益率负相关。IC 绝对值越大，表明因子对收益率贡献率越高。通过查阅文献，本文设定 IC mean>0.025 的因子较为有效。

IC 值可通过计算全部股票在调仓周期期初排名和调仓周期期末收益排名的线性相关度

（Correlation)得出：

$$IC^T=correlation(r^{T+1},X^T)$$

其中，IC^T表示因子在 T 期的值，r^{T+1}表示第 T+1 期的收益率，X^T表示第 T 期因子 X 的暴露度。

IR 即信息比率，是超额收益率的均值和标准差之比，表示因子稳定获取超额收益的能力，可判断出基于 IC 判断的因子选股稳定性，因子 IR 越大， 代表因子在历史上表现的稳定性。通过查阅文献，本文设定 IR>0.3 的因子较为有效。

IR 可以通过 IC 近似计算得出：

$$IR\approx\frac{\overline{IC_t}}{std(IC_t)}$$

因子在不同的历史时期的表现有可能差别很大.，表现在 IC 上，就是 IC 的波动率很大.。假设 IC 均值一定,，IC 的波动率越小，表现越稳定,，IR 就越大。

单因子 IC 分析法是多因子模型中常用的检验因子有效性的方法。因为 IC 允许我们对因子的预测能力做出评价。在研究因子的表现时，只能采取历史的数据进行评估，这也就意味着无法保证未来的组合收益率能够跑赢市场。而 IC 通过链接本其因子值与下期股票收益率，推断出未来的收益率。

而 IR 是 IC 的多周期均值与 IC 的标准方差，通过将一段时间内 IC 的均值纳入计算，能够显示出因子选股的稳定性，避免选取那些仅在短时间内显示出与股票收益率具有相关性的因子。

基于上，本文将选择题干中提供的三十支股票从 2015 年 1 月 1 日到 2020 年 12 月 31 日的数据为样本，对备选因子库的中的所有因子进行单因子 IC 检测。

细分因子	IC 均值	IR
MKV	0.0071	9.09%
NEG_MKV	-0.0103	-11.75%
PB	0.0076	9.99%
PS	0.0088	11.89%
PE	0.0089	12.96%
ROE	0.0045	6.27%

ROA	0.0052	6.78%

单因子 IC、IR 值统计表

5.2.2 因子相关性检验：相关系数的计算

在对备选因子进行了有效性的检验之后，容易发现，经过有效性初步筛选的备选因子有些属于同一大类因子，如市净率（PB）、市销率（PS）、市盈率（PE）同属于价值类因子。这些备选因子背后的作用逻辑较为相似，而如果把这些因子都纳入模型中进行考虑，很有可能加剧了这类因子的叠加作用。从而使得这大类因子对整个选股策略产生更大的影响，降低其他大类因子的作用。因此本文对初步筛选后的因子进行相关性检验，具体方法如下：

各股票的在所选时段（2015 年 1 月 1 日至 2020 年 12 月 31 日）的因子值和有效因子可以构成一个 n × k 的矩阵（ n 行代表因子值， k 列代表各股在时段内的因子值）。之后对该矩阵的所有列进行相关系数的求解，以得到各股因子值在不同因子下的相关性矩阵。

图表

文献研究表明，因子值相关系数阈值ＭｉｎＣｏｒｒ应取 0.5，一旦两个因子之间的相关性系数小于这个阈值，就可认为着两个因子之间具有较强的独立性。而如果两个因子之间的相关性系数大于这个值，我们即可认为它们之间存在较大的相关性，是同类型因子，本文将在下一环节对这些同类因子进行合并。

	总市值	**流通市值**	**市净率 PB**	**市销率 PS**	**市盈率 PE**	**权益回报率 ROE**	**资产回报率 ROA**
总市值	1	0.897608	-0.11091	-0.51759	-0.46946	0.728772	-0.03994
流通市值	0.897608	1	-0.26813	-0.62253	-0.61702	0.770432	-0.08066
市净率 PB	-0.11091	-0.26813	1	0.831149	0.828768	-0.11892	0.570352

市销率 PS	-0.51759	-0.62253	0.831149	1	0.919312	-0.42994	0.431751
市盈率 PE	-0.46946	-0.61702	0.828768	0.919312	1	-0.53835	0.329154
权益回报率 ROE	0.728772	0.770432	-0.11892	-0.42994	-0.53835	1	0.24381
资产回报率 ROA	-0.03994	-0.08066	0.570352	0.431751	0.329154	0.24381	1

因子间相关性矩阵

经过检验后我们发现，同一类型的细分因子之间的相关性较高。规模类因子下的总市值和流通市值的相关性较高，达到 0.897608。市净率、市销率、市盈率这三个价值类因子的细分因子之间的相关性均大于 0.8，相关性较强。因此，本文判断在同一大类因子下的细分因子之间有较高相关性，后续拟将相关性较高的因子进行合并。

5.2.3 相关因子的合成：主成分回归法

通过上述计算，我们得到了因子之间的相关系数，并计划对相关性较高的因子进行合并。在很多研究中，对于相关性较高的因子都是采取保留表现较好的因子，而舍弃表现较差因子的方法。本文认为，这种方法没有缩窄了所考虑数据的宽度，同时将所有的风险都暴露在某一个因子下，这会导致因子的选股能力降低和风险的增加。于是，本文将相关性较高的因子进行合并，增加所考虑的数据，以求在最大化各因子有效性带来的溢价同时降低单个因子的风险暴露。本文采取两种方法对细分因子进行合成：

（1）等权法

（2）PCA 方法（主成分分析法）

本文利用 PCA 法降维和去除相关性的功能，来解决同类细分因子维数过高，以及细分因子间的相关性给多因子选股带来的困难。

PCA 的做法是，对 X 作正交变换，寻求原指标的线性组合Y_I。

$$\begin{cases} Y_1 = b_{11}X_1 + b_{21}X_2 + \cdots + b_{p1}X_p \\ Y_2 = b_{12}X_1 + b_{22}X_2 + \cdots + b_{p2}X_p \\ \cdots \\ Y_p = b_{1p}X_1 + b_{2p}X_2 + \cdots + b_{pp}X_p \end{cases}$$

满足如下的条件：

每个主成分的系数平方和为 1，即

$$b_{1i}^2 + b_{2i}^2 + \cdots + b_{pi}^2$$

主成分之间相互独立，即

$$cov(Y_i, Y_j) = 0, \qquad i \neq j, \qquad i, j = 1,2,3, \ldots, p$$

主成分的方差依次递减，重要性依次递减，即

$$Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p)$$

基于以上条件确定的综合指标Y_1，Y_2，…，Y_p分别称为原始指标的第一个主成分，第二个主成分，……第 p 个主成分。其中，各综合指标在总方差中所占的比重依次递减，在实际研究中，本文挑选前几个方差最大的主成分。

主成分对收益率的贡献率

指标	$\boldsymbol{\rho}$
PB	90.2564
PS	6.3767
PE	3.3669

这里我们主要根据主成分对收益率的贡献度来分配因子的权重。

$$\omega_{IC}^{i} = \frac{\boldsymbol{\rho_i}}{\sum_{i=1}^{3} \boldsymbol{\rho_i}}$$

其中，ω_{IC}^{i}表示第 i 个细分因子在大类因子中分配的权重，**$\boldsymbol{\rho_i}$表示第 i 个细分因子的贡献率**

5.2.4 大类因子的权重分析： IC 加权法合成

不同的大类因子之间其因子有效性是存在差异的，根据以往的研究结果，一般而言估值因子和规模因子都是表现相对显著的，而杠杆，运营因子都是表现相对较差的因子，如果以等权的方式来对各大类因子进行加权，则忽略了不同因子的解释力度。故本文大类因子的 IC

均值加权合成。大类因子的 IC 均值加权配置，即根据前几期大类因子的 RANK IC 的均值对当期的大类因子配置权重。此次研究我们将大类因子在过去 5 年的 IC 均值作为新大类因子的权重。

IC 加权法的具体步骤：

1.计算第 t 大类因子的平均 IC：

$$\frac{1}{n}\sum_{i=1}^{n} \mid \mathrm{IC}_{mean}^{i} \mid = \overline{\mathrm{IC}_{mean}^{t}}$$

其中，IC_{mean}^{t}表示第 t 大类因子的平均 IC，IC_{mean}^{i}表示第 t 大类下 n 个细分因子的 IC 均值。

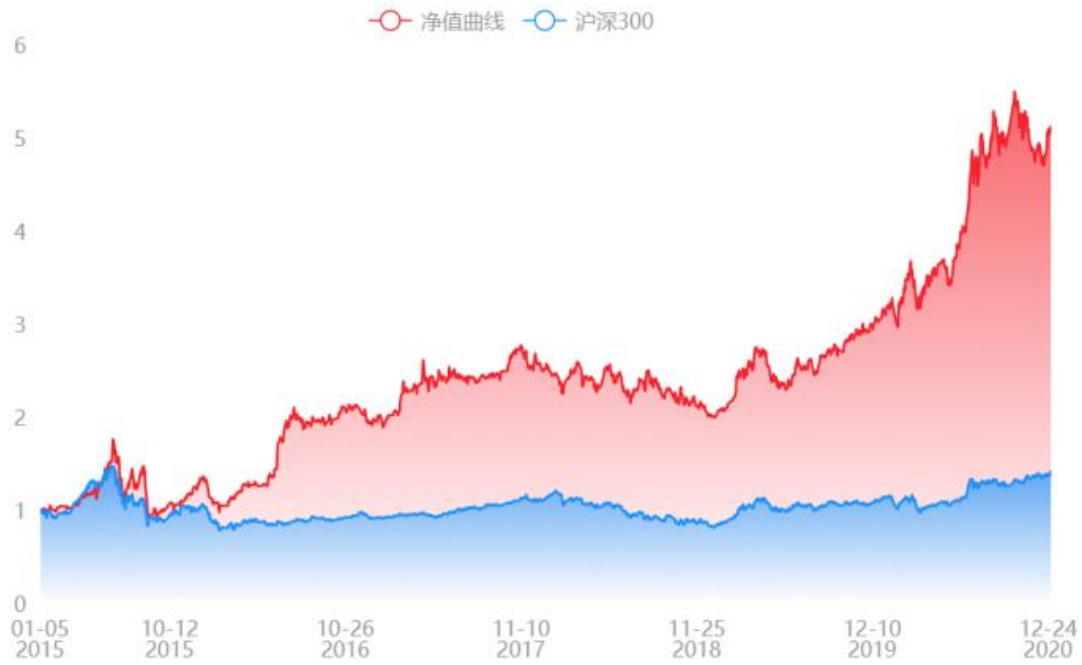
2.分配 3 个大类的权重：

$$\omega_{IC}^{t} = \frac{\overline{\mathrm{IC}_{mean}^{t}}}{\sum_{t=1}^{3} \overline{\mathrm{IC}_{mean}^{t}}}$$

其中，ω_{IC}^{t}表示第 t 大类因子的权重。

5.2.5 模型选股效果

基于上述步骤构建起的模型，本文从中挑选出表现最好的六支股票作为选股组合，组合内部各个股的投资资金量采用等权法，使用 AT 平台对其进行回测，得到结果如下图。



IC 加权法下模型收益曲线与市场指数收益曲线

回测结果显示模型选股累计收益达到 413.45%，年化收益率达到 32.58%，在绝大多数情况下能够跑赢市场指数，表现良好。

同时，为了比较不同因子合成方法对收益率的影响，本文在不改变其他条件的情况下，对大类因子权重采取等权法进行分配。使用 AT 平台对其回测的结果如下图。

等权法下模型收益曲线与市场指数收益曲线

建表对两模型的选股能力和各评价标准进行比较：

策略名称	年化收益率	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤	换手率	回测时间
等权策略	29.97%	43.11%	0.22	**0.85**	**0.95**	**1.08**	**44.40%**	**489.99%**	2015-01-05~2020-12- 31
IC 加权策略	32.58%	43.11%	0.24	0.86	1.03	1.21	48.54%	374.48%	2015-01-05~2020-12- 31

5.3 问题三：外部环境对模型内因子权重的影响

在股票市场上，投资者们往往通过量化模型来估计当前市场的情况和预判未来市场的趋势，本文介绍的多因子模型就是量化模型中的一种。然而，传统的量化模型只能根据股票某些典型特征进行分析预测，无法考虑外部因素的影响，因此，本文根据研究文献探索题干给出的三种外部因素（突发事件闪现、舆情影响、自然灾害影响）会如何影响上文中建立的模型。

5.3.1 突发事件闪现

突发事件之所以被称为突发事件，根本的原因是发生不具有可预测性。例如（宝能系旗下的钜盛华通过资管计划，在 2015 年 11 月 27 日至 12 月 4 日期间大举买入 5.49 亿股万科 A，对万科 A 股价的波动影响。突然事件的闪现往往会在极短的时间内对股票市场产生极大的冲击，使股市收益出现剧烈波动。但研究表明，在突发事件发生后，后验地考察股票市场发生的变动，就能够从中观察出突发事件对股票的影响。

5.3.2 舆情影响

从定义上来说，舆情对股票市场影响是指由某些事件引发的投资者态度、情感和行为倾向在股票价格上的反映。例如 2018 年世界杯，华帝公司启动“法国队夺冠华帝退全款”活动，华帝 股份的舆情营销策略，公司股价 7 月 2 日却跌停。而随着互联网的不断发展，线上投资得到普及，各股票交易平台也已经发展完善。这也意味着，公司舆情信息能够更快更全面的反映到股票中去，因此，在多因子模型之中考虑舆情对股市的影响势在必行。

5.3.3 自然环境影响

自然环境的影响主要指的是自然灾害对人类社会造成的冲击与伤害。例如 2008 年的四川汶川大地震、2010 年青海玉树地震、2020 年新冠疫情的爆发等等……相关研究表明，自然灾害会通过影响上市公司的基本面来营销股票收益。一般认为有两种作用逻辑：一是自然灾害损害了公司财产，导致基本盘收到影响，从而冲击该公司股票；二是投资者情绪、行为受到自然灾害的影响，是行为金融学研究的范畴。

5.3.4 综合分析

综合上诉细分外界环境影响因素，本文发现，外部环境因素主要的影响逻辑为改变市场大环境，不同的事件会推动投资者对市场的判断向不同方向变化。而本文认为，对市场大环境的影响可以通过改变规模类因子、成长类因子这两类反映公司稳定性和成长性的因子的权重反映在模型中。

在外部环境因素影响市场环境向积极发展的时候，我们认为投资者会更多的考虑投资的超额回报率，而不是投资收益的稳定性。因此，此时我们降低规模类因子的权重，增加质量类因子的权重。

而当外部环境因素影响市场向消极方向发展时，我们认为投资者会看重投资的稳定性，而将投资的超额回报率放在第二位。因此，此时我们降低质量类因子的权重，提升规模类因子的权重。

	市场总体向好	**市场总体变差**
规模类因子	权重下调	权重上调

质量类因子	权重上调	权重下调

外部影响因素在因子权重上的反映

5.4 问题四：考虑外部环境影响的的多因子选股模型搭建

根据问题三的分析，并将改进后模型在观测期（2015 年 1 月 1 日到 2020 年 12 月 31 日）进行回测，从而计算出在该模型下收益率最高的十支股票作为本文最终的选股投资策略。

考虑外部影响因素时模型收益曲线与市场指数收益曲线

策略名称	年化收益率	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤	换手率	回测时间
考虑外部影响	6.38%	25.51%	-0.06	0.46	0.24	-1.08	14.63%	2019.10%	2020-01-02~2020-12- 31

6 模型的评价与改进

本文选取了质量因子、规模因子、价值因子三个方面共 7 个因子，并对其进行有效性检验和相关性检验，证明了这些因子的变动与超额收益率之间的相关性。同时，根据这 7 个有效因子，利用 PCA 法（主成分分析法）和 IC 加权法构建模型，本文得以从题干三十支股票中选出十支表现最好的股票，在回测期（2015 年 1 月 1 日到 2020 年 12 月 31 日）中的累计收益达到 413.45%，年化收益达到 32.58%，表现良好。

其后，通过对外部因素对模型因子的影响的评估，将舆情影响、突发事件的闪现和自然环境影响纳入考虑。

6.1 模型的优点与缺点

优点：

1.采用了 PCA、IC 加权等多种方法，将小类因子合并，对大类因子分配权重，即增加了信息的有效利用率，又提高了最终模型的合理性。

2.备选因子库的选取基于多方面的考量，在定性分析上也提高了多因子选股模型的收益率。

缺点：

1.因子有效性的判断指标较少，可信度不高。

2.相关性较高的因子进行合并时采用的方法过于单一，没有考虑各大类型因子的特性。

3.没有将外部影响因素提取为因子纳入模型进行考虑，评价方式较为主观。

6.2 模型的改进方向

在判断因子有效性的环节，本文仅选取了 IC 均值与 IR 值作为评价因子是否有效的指标。即使上，研究显示，如将因子 IC>0 的比例、IC 绝对值>0.02 的比例等判断指标纳入考虑，对因子有效率的检测会更加有可信度。还可以引入机器学习的方法探索检验模型有效性的新方式。

在检验因子相关性后，本文采取了因子合并的方式降低同类型因子对模型的叠加作用，

但在细分因子合并环节只采取了单一的 PCA 法（主成分回归法）。而先前研究发现，等权法、PCA 法、逐步回归法这三个常见的因子合并方法在不同的大类因子合并中各有优劣。所以，要使模型的选股能力得到提高，应对不同大类的细分因子采用不同的合成方法。

在考虑外部影响因素的环节，本文仅通过文献研究的方式大致确定了其对模型内因子的影响，粗略调整了投资策略。未来应该通过机器学习的方法，将这些非结构化的数据转化为时序数据，以便将其构建成特征因子加入模型进行考虑。

7 参考文献

[1]林晓明，华泰多因子模型体系初探，华泰证券研究报告，2016

[2]陈栌，基于多因子选股模型的 A 股投资策略，浙江工商大学，2016

[3]许闲；刘淇；王怿丹，自然灾害的股价效应——来自 A 股市场的实证证据，世界经济文汇，2021

[4] 耿素娟；张莉，大数据背景下网络舆情对股票收益率的影响研究，全国流通经济，2021

[5] 邓琪;于跃，突发事件对中国股市的冲击效应，统计与管理，2021

[6] 江方敏，基于多因子量化模型的 A 股投资组合选股分析，西南交通大学，2013

[7] 雷璇，基于回归法和打分法的因子选股模型对比分析，大连理工大学，2019

8 附录

数据处理（python）

```
import pandas as pd
df=pd.read_excel(r'factor_data.xlsx')
#标准化
for i in range(df.shape[1]):
    mean_f = df.iloc[:,i].mean()
    std_f = df.iloc[:,i].std()
    df.iloc[:,i] = (df.iloc[:,i]-mean_f)/std_f
#去极值（3σ 法，又称为标准差法）
for i in range(df.shape[1]):
```

```
        median_f = df.iloc[:,i].median()
        new_median_f = ((df.iloc[:,i]-median_f).abs()).median()
        max_f = median_f + 3*new_median_f
        min_f = median_f - 3*new_median_f
        for j in range(df.shape[0]):
            if df.iloc[j,i] > max_f:
                df.iloc[j,i] = max_f
            if df.iloc[j,i] < min_f:
                df.iloc[j,i] = min_f
            else:
                continue
df.to_excel(r'factor_data_processed.xlsx')
```

PCA(matlab)

```
clc,clear
data=xlsread('质量类.xlsx'); %读取的数据，一行为一组
%data=zscore(data); %zscore 标准化
r=corrcoef(data); %计算相关系数矩阵
%利用 r 进行主成分分析，x 的列为 r 的特征向量，即主成分的系数
[x,y,z]=pcacov(r); %输出贡献率及其它结果
f=repmat(sign(sum(x)),size(x,1),1);
x=x.*f;
num=2; %选取的主成分数目（根据贡献率来定）
df=data*x(:,[1:num]);
tf=df*z(1:num)/100;
[stf,ind]=sort(tf,'descend');
%result=[ind,stf] %输出序号和综合得分（降序）
%求载荷矩阵
[vec,val,con]=pcacov(r);
num=3;
f1=repmat(sign(sum(vec)),size(vec,1),1);
```

```
vec=vec.*f1;

f2=repmat(sqrt(val)',size(vec,1),1);

a=vec.*f2;

aa=a(:,1:num);%载荷矩阵

s1=sum(aa.^2);

s2=sum(aa.^2,2);
```

策略代码（matlab）

```
function pca_ic_Stratey(bInit, bDayBegin, cellPar)


len = cellPar{1};

len1=cellPar{2};

len2=cellPar{3};

num=cellPar{4};

% bInit: bool, is init action

% bDayBegin: bool, is bar day begin

% cellPar: cell, run mode function's third param, params transaction used

global factor1;

global factor2;

global factor3;

global factor4;

global factor5;

global factor6;

global factor7;

global g_idxKweek;

global TLen;


global Tlen;


if bInit

    traderSetParaIMode(false);
```

```
% Initialization operation
% Register data,
g_idxKweek = traderRegKData('day', 1);
% Get Registered data, e.g:
%weight9 = traderGetRegKData(g_idxCCfx,1);
TLen = length(g_idxKweek(:,1));
%, bpPFCell Optional
% g_idxK traderRegKData return value,when call traderRegUserIndi may need it, post as
first param
% Registration factor, e.g:
%g_idxCCfx = traderRegUserIndi(@g_idxccfx, {g_idxKweek,len,len1,len2,num})
g_idxKweek = traderRegKData('day',1);        %注册 K 线数据
%注册因子数据

factor1 = traderRegFactor('mkv');    %总市值
factor2 = traderRegFactor('neg_mkv');    %流通总市值
factor3 = traderRegFactor('PB');        %市净率
factor4 = traderRegFactor('PS');        %市销率
factor5 = traderRegFactor('PE');        %市盈率
factor6 = traderRegFactor('roe');        %权益回报率
factor7 = traderRegFactor('roa');        %资产回报率

% TODO Write your initialization operation
else
% TODO Write your strategy code
% targetList = traderGetTargetList();  %获取标的
Tlen = 10;  %计算标的个数
%小类因子数据读取与处理
factor1_data = traderGetRegFactor(factor1,1);%获取因子数据
factor1_winsorized = winsorize(factor1_data,[5,95]); %极值处理
```

```
factor1_standardized = standardize(factor1_winsorized);  %标准化

factor2_data = traderGetRegFactor(factor2,1);%获取因子数据
factor2_winsorized = winsorize(factor2_data.^(-1),[5,95]); %极值处理
factor2_standardized = standardize(factor2_winsorized);  %标准化

factor3_data = traderGetRegFactor(factor3,1);%获取因子数据
factor3_winsorized = winsorize(factor3_data.^(-1),[5,95]); %极值处理
factor3_standardized = standardize(factor3_winsorized);  %标准化

factor4_data = traderGetRegFactor(factor4,1);%获取因子数据
factor4_winsorized = winsorize(factor4_data.^(-1),[5,95]); %极值处理
factor4_standardized = standardize(factor4_winsorized);  %标准化

factor5_data = traderGetRegFactor(factor5,1);%获取因子数据
factor5_winsorized = winsorize(factor5_data.^(-1),[5,95]); %极值处理
factor5_standardized = standardize(factor5_winsorized);  %标准化

factor6_data = traderGetRegFactor(factor6,1);%获取因子数据
factor6_winsorized = winsorize(factor6_data.^(-1),[5,95]); %极值处理
factor6_standardized = standardize(factor6_winsorized);  %标准化

factor7_data = traderGetRegFactor(factor7,1);%获取因子数据
factor7_winsorized = winsorize(factor7_data.^(-1),[5,95]); %极值处理
factor7_standardized = standardize(factor7_winsorized);  %标准化

%小类因子合并
res_1 = 0.5*factor1_standardized + 0.5*factor2_standardized;
res_2 = 0.9*factor3_standardized + 0.064*factor4_standardized +
0.036*factor5_standardized;
```

```
res_3 = 0.5*factor6_standardized + 0.5*factor7_standardized;
%大类因子权重
score_weight = [0.3936,0.3826,0.2211];
%score_weight = [1/3,1/3,1/3];
total_score = res_1 * score_weight(1) + res_2 * score_weight(2) + res_3 * score_weight(3); %算因子加权后总分

[~,I] = sort(total_score,'descend');
[~, MarketCap,~, ~,~] = traderGetAccountInfoV2(1);  %获取最新的权益

SelectedID = I(1:num); % 选出 alpha 最大的一组股票
Stock_flow = ((MarketCap)*1)/num; % 每只股票分配的资金
stock_list=1:num;%change
a1=ismember(stock_list,SelectedID);



dataDay = traderGetRegKData(g_idxKweek,1,false);  %获取最新的价格数量数据
latest_close = dataDay(5:8:end);  %取得收盘价数据

%target_position = zeros(Tlen,1); %初始化目标仓位
%target_position (index_selcted) = HandListCap * weight;  %得到最新一起目标仓位
target_position = zeros(Tlen,1); %初始化目标仓位
target_position (a1) = Stock_flow; %得到最新一起目标仓位


mp = traderGetAccountPositionV2(1,1:Tlen);
%%
%接下来是调仓部分
```

```
    for i = 1:Tlen
        num=floor(target_position(i)/latest_close(i)/100)*500;
        traderPositionToV2(1,i,num,0,'market','rebalance');
    end
end
end

function y = standardize(x)
% STANDARDIZE    standardize a vector
% INPUTS         : x - n*1 data vector
% OUTPUTS        : y - winsorized x, n*1 vector
if ~isvector(x)
    error('Input argument "x" must be a vector')
end
y = (x-mean(x,'omitnan'))/std(x,'omitnan');
end

function [y,varargout] = winsorize(x,p)
% WINSOR         winsorize a vector
% INPUTS     : x - n*1 data vector
%              p - 2*1 vector of cut-off percentiles (left, right)
% OUTPUTS    : y - winsorized x, n*1 vector
%              i - (optional) n*1 value-replaced-indicator vector
% NOTES      : Let p1 = prctile(x,p(1)), p2 = prctile(x,p(2)). (Note
%              that PRCTILE ignores NaN values). Then
%              if x(i) < p1, y(i) = min(x(j) | x(j) >= p1)
%              if x(i) > p2, y(i) = max(x(j) | x(j) <= p2)

if ~isvector(x)
    error('Input argument "x" must be a vector')
```

```
end
if nargin < 2
    error('Input argument "p" is undefined')
end
if ~isvector(p)
    error('Input argument "p" must be a vector')
end
if length(p) ~= 2
    error('Input argument "p" must be a 2*1 vector')
end
if p(1) < 0 || p(1) > 100
    error('Left cut-off percentile is out of [0,100] range')
end
if p(2) < 0 || p(2) > 100
    error('Right cut-off percentile is out of [0,100] range')
end
if p(1) > p(2)
    error('Left cut-off percentile exceeds right cut-off percentile')
end
p = prctile(x,p);
i1 = x < p(1); v1 = min(x(~i1));
i2 = x > p(2); v2 = max(x(~i2));
y = x;
y(i1) = v1;
y(i2) = v2;
if nargout > 1
    varargout(1) = {i1 | i2};
end
end
```