金融计量
——基于公司关联关系的图神经网络股票收益率预测

朴泰俊 詹天宇 卢浩楠

北京大学国家发展研究院
National School of Development

摘　要

股票收益率预测一直是金融学界和业内研究的重点课题。然而，主要的研究方法与模型工具大都聚焦于通过个股的价格、基本面数据进行截面或时序预测，表征股票间关联关系的一些另类基本面数据的应用——尤其在中国股票市场——则鲜少有人涉猎。

本文将目光聚焦于上下游产业间的供求关系，利用了中国 A 股上市公司间表征公司供求关系的供应链基本面数据，自主构造了供应链关系矩阵，基于不同的传统价量因子生成了供应链传导因子并检验了其有效性。之后利用由 LSTM 时序循环神经网络和 TGC 时态图卷积神经网络组成的 RSR 模型，综合学习了股票价格基本面数据和供应链关系数据，训练出了最终的深度学习因子，并进行五分组等权回测，获得了稳健良好的回测效果。

关键词：供应链传导，LSTM，图卷积神经网络，收益率预测

1 引言

1.1 基于公司间关联的收益率预测

二十世纪七十年代，实证资产定价（Empirical Asset Pricing）作为理论资产定价的补充开始兴起。其主要目的是以真实发生的市场数据为基础，检验已有资产定价理论，解释金融异象。基于 CAPM 模型，Ross（1976）提出了套利定价理论（Arbitrage Pricing Theory，APT），套利定价认为，如果市场未与达到均衡状态的话，市场上就会存在无风险套利机会，获得稳定的超额收益，并且可以用多个因子来解释风险资产收益。随后，诺贝尔经济学奖获得者 Fama 进一步提出了三因子模型（Fama and French，1993）和五因子模型（Fama and French，2015）等。这些实证资产定价模型刻画了各类因子与市场预期收益之间的关系，为理解市场金融异象和市场波动提供了有效的理论依据和方法指导 [谭 22]。

然而，不论是单因子 CAPM 模型，还是 Fama-French 三因子等多因子模型，这些资产定价模型都是从企业自身的视角出发，以单一资产的视角，解释不同风险因素如何影响股票资产的预期收益问题，并未考虑了资产之间的关联关系对不同资产预期收益率的影响作用。事实上，资本市场是一个涵盖多样资本的，复杂且不断变化的动态市场。在市场中，各个股票资产因其内在价值的关联性、公司之间的合作与竞争、投资者对不同资产的认知和比较、以及监管部门的监管需求而产生不同类型的关联，构成了一个复杂的动态关联公司网络。在这个企业关联网络中，某一企业的状况，无疑会对其关联的企业在市场中的表现产生不同程度的影响。

大量金融研究发现，基于某种联系关联起来的企业是影响资产价格波动的重要因素，这种影响经常表现为股票收益率的“领先—滞后效应 (lead-lag effect)”。例如，Moskowitz and Grinblat(1999)[MG99] 首先提出了行业动量的存在，他们发现过去行业回报可以预测未来股票回报，即使在控制了个股层面动量后也依然显著。Rashes（2001）[Ras01] 发现只有股票名称相似，几乎没有其它共同点的公司之间，股票收益存在显著的正相关性。Parsons et al.（2020）[PST20] 发现了区域动量的存在，指出总部位于同一地理区域的公司的股票，未来收益率之间具有显著的相互影响作用。

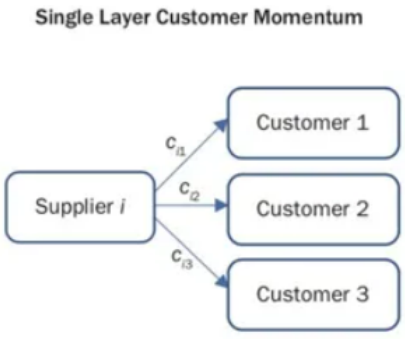

图 1: 一级供应链关系

而在这些成果中，有一篇文章引起了作者的强烈兴趣。Lauren Cohen 与 Andrea Frazzini

两位著名学者 2008 年在 The Journalof Finance 发表了一篇的文章: Economic Links and Predictable Returns[CF08] 在当时引发了广泛讨论与深远影响。这篇文章发现，上市公司间的关联性以及股价波动会沿着公司间的**供应链**传导，即过去客户的回报可以预测未来供应商公司的回报。例如当某一上游企业因为丑闻导致其股价下跌时，其下游的客户企业股价很可能收到波及而同样下跌。Cohen and Frazzini 基于每个公司对应的供应商和客户集合定义了公司间的经济关联，例如图1为一个一级供应商-客户关系示意图，当市场参与者没有充分意识到这种关联性的存在，即市场并不完全有效，那么相关联公司间的股价就会出现 Lead-Lag 效应。基于此开发的因子会有不错的收益。

因此，如何在资产定价模型中考虑公司之间复杂多样的关联关系，量化相关联股票资产之间的溢出影响，是理解影响股票资产预期收益本质规律，实现对复杂市场运行过程合理建模的一个重要研究方向。

这些经济关联的文献运用美国的数据，从实证的角度验证了市场并非完全有效。然而中国股票市场在发展历程、投资者构成、市场有效性方面与美国股票市场有较大的差异，各类经济关联因子在中国股票市场上是否具有显著的预测能力是一个有待回答的实证问题。一方面，中国股票市场换手率高 [Pan t a.2016)[PTX15] 股化信息快，经济关联企业的信息有可能更快地反映到标企/的股价中，另一方面，中国股票市场起步较晚、市场有双性相对较低，并且散户较多，投资者注意力和信息分析能力更加有限，这可能导致投资者对企业间的关联识别效率较低，经济关联的信息融入股价速度更缓慢，经济关联因子预测能力更显著。因此，本文对各类经济关联因子在中国股票市场上的特征进行初步探究，以期推进对中国股票市场定价规律的理解。部分学者已经关注到了中国股票市场中的公司关联，胡聪慧等 (2015) 利用行业分部的

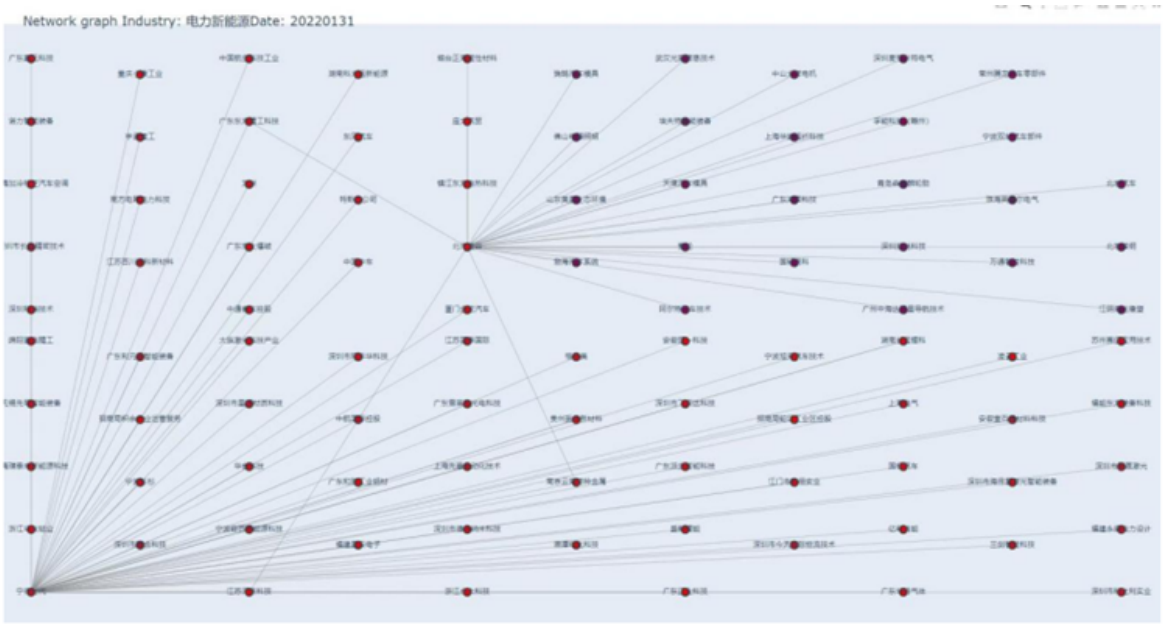

图 2: A 股市场供应链关系：以宁德时代为例来源：中信证券研报

信息，实证发现单一行业公司过去的股票回报可以在月度层面预测集团公司未来的股票回报，

并提出这一预测性源自投资者的有限注意 [张 22]。然而，前所提到的供应链关联、区域关联、科技关联等经济关联还均未在中国股票市场中得到验证。本文试图从供应链的角度出发，尝试探究中国股票市场中供应商与客户关系所带来的附加预测能力与预测周期，以期补充这一领域的研究。

图2为与宁德时代有关的上下游供应链关系网络。

1.2 股票时序数据的建模方法

股票时序数据的传统建模方法基于时间序列分析，例如卡尔曼滤波、自回归模型及其拓展模型。对于股票的某一个量化特征，比如价格、收益率等，这些模型将其视为一个随机过程，使用历史数据拟合模型参数。

这些主流模型有其不足之处：

- 经常包含平稳性，正态性等模型假设，而实证数据往往不满足这些假设
- 模型参数随着量化指标的增加和数据量的增加而急剧增大，拟合难度也迅速上升

因此这些模型尚不足以完美地刻画受多种因素影响的，变化模式较为复杂的股票市场。针对这些缺点，一些更加高级的模型逐渐被开发并用来代替传统时序模型预测未来股价的走势。其中，专门用来处理序列数据的递归神经网络 (RNNs) 被认为大有前景。目前较为先进的递归神经网络序列模型是 “状态频率内存网络 (SFM)”，它着眼于挖掘不同频率下的短期股价结构，并比较成功地预测了美股日开盘价。

不过，SFM 仍有可以改进之处。典型机器学习模型总是尽可能尝试精确预测股票涨跌方向或者最小化均方误差/绝对误差，但是最小的损失并不一定导致最优的选股策略和最佳净值曲线，即传统的对于预测值和真值的误差计算方法在选股投资语境下并不是最优的选择。此外，结合上一小节所述的，股票之间的关联性会使得股票收益率相互影响，过去的股票关联信息包含未来股票的收益信息。而前述的神经网络模型仍仅将所有股票分开处理，分别对每个时序数据建模，从而忽略了公司之间丰富的关联。

为了将股票关联与收益预测整合在一起，使用图结构代表股票关联是一种直观的想法。Fuli Feng and Xiangnan He(2019)[FHW+19] 提出了一种基于时态图卷积神经网络 (Temporal Graph Convolution) 的训练方法。将经过时序处理股票量价数据和股票关联数据 (wiki relation) 通过嵌入 (embedding) 得到一个三维表征向量，再将其输入 TGC 网络进行训练拟合。将该方法应用于美国纳斯达克/纽约股票交易所的股票数据后，可以获得超出基准方法 (SFM) 超过 115% 的收益净值。

1.3 本文工作

受 Fuli Feng and Xiangnan He[FHW+19] 文章的启发，再结合前述探究股票关联关系对股票收益预测的想法，本文试图在 A 股市场应用基于公司供应链关联关系的图神经网络股票收

益率预测。首先，我们使用 python 搭建股票多分组选股回测框架，并收集了 A 股公司上下游供应链关系，生成供应链矩阵；再以传统动量因子作为基准，展示了经过供应链关系传导的动量因子包含了额外的对股价的预测信息，证实了股票关联关系可以预测股票收益；之后，通过改进供应链矩阵的定义和合成方法，以及更精细化地处理动量传导的形式，得到了更好的回测绩效；最后，我们引入深度学习方法，先将包含量价特征的股票时序数据输入长短期记忆递归神经网络 (LSTM)，再将该输出和供应链矩阵数据进行嵌入 (embedding) 操作，合成一个综合表征向量以后输入图卷积神经网络训练，最终得到可供排序选股的网络输出结果，进行多分组选股回测。此外，我们定义了不同于传统损失函数的，基于股票排序关系的新损失函数，使得模型训练过程更富经济学含义。

本文结构安排如下：第 2 部分介绍研究方法，包括因子回测流程，供应链矩阵的生成和传导，神经网络的构建和原理，损失函数的选取；第 3 部分介绍我们所用的 A 股市场量价数据和供应链数据；第 4 部分介绍实证研究过程和最终结果；最后是结论与展望。

2 研究方法

2.1 股票关联数据对于因子的增强

本部分阐述如何将股票关联关系应用于收益率预测。

顾名思义，关联关系指代股票之间两两成对出现的一种指标，其在数学上的具体表现往往是矩阵形式。如下所示，矩阵 A 为我们收集的 A 股供应链关系数据 (详见第 3 部分数据描述) 的典型矩阵表述

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

其中矩阵的行指标 i 为 A 股上市公司 (股票)，列指标 j 为 (1) 自然人或 (2) 机构。a_{ij} 表示第 j 个下游的自然人/机构对 (我们关心的) 第 i 个股票的影响，且总是有 $n > m$, 即行只包括了 A 股，而列包括了未上市机构、A、B、港、美股等上市的公司，以及自然人。

构建矩阵的数据包含多种形式的供应链关系，例如一级关联、二级关联、母子公司关联等，数据的具体描述请见第 3 部分。

常见的传统选股因子是基于某支股票过去一段时间的基本面或量价数据经过一定方式的计算得到，具有 M_{jt} 的形式，表示 t 时第 j 支股票的因子暴露度。因此，为了将股票关联的信息引入已经被证明有效的因子中，我们将关联矩阵与因子矩阵相乘，得到 t 时第 i 支股票的供应链传导因子为

$$F_{it} = T_{ij} M_{jt}$$

为了证明其有效性，我们再将新因子对原因子线性回归，取残差进行回测检验。

2.2 因子回测方法

按照统计套利方法的标准流程进行有效因子识别与回测：

1. 单因子回归确定每期因子收益

$$\widetilde{r}_j^t = \sum_{s=1}^{s} X_{js}^t * \widetilde{f}_s^t + X_{jk}^t * \widetilde{f}_k^t + \widetilde{u}_j^t$$

$\widetilde{r}_j^t$：股票 j 在第 t 期的收益率
X_{js}^t：股票 j 在第 t 期在行业 s 上的暴露 (一个 0-1 哑变量，表示股票是否属于该行业)
$\widetilde{f}_s^t$：行业 s 在第 t 期的收益率
X_{jk}^t：股票 j 在第 t 期在因子 k 上的暴露
$\widetilde{f}_k^t$：因子 k 在第 t 期的收益率

每日以收益率为因变量对因子暴露做线性回归，可得到 $\widetilde{f}_k^t$ 序列和累积收益率序列

2. IC 值

因子的 IC 值是指个股第 t 期在因子上的暴露度与 t+1 期的收益率的皮尔逊相关系数。因子 IC 值反映的是个股下期收益率和本期因子暴露度的线性相关程度，是使用该因子进行收益率预测的稳健性。使用 corr() 函数直接计算 IC 值，可得到 IC 序列及其统计指标。

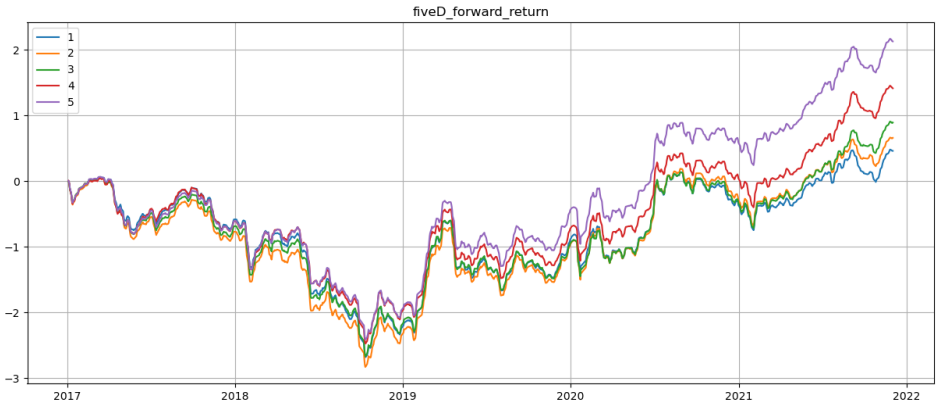

图 3: 五分组因子回测示意图

3. 多分组回测

在每个截面期 (本文为日频) 上，根据因子值对个股进行打分，将所有个股依照分数进行排序，然后分为 N 个投资组合，将每个投资组合内部的股票等权做多，根据下一期股票的收益率计算每个投资组合的当日净回报。累加该回报便得到了多分组的回测净值曲线，图3是一个 $N = 5$ 的回测曲线示意图。

评价回测结果的指标：回测年化收益率、年化波动率、夏普比率、最大回撤、胜率等。

2.3 时序图卷积神经网络方法

参考 Fuli Feng and Xiangnan He(2019)[FHW+19] 的文章，我们的模型构建如图4所示，该模型包含三个层次，分别是时序嵌入层 (下面部分)，关系嵌入层 (中间部分) 和一个预测层 (上面部分)，下文将进行详述。

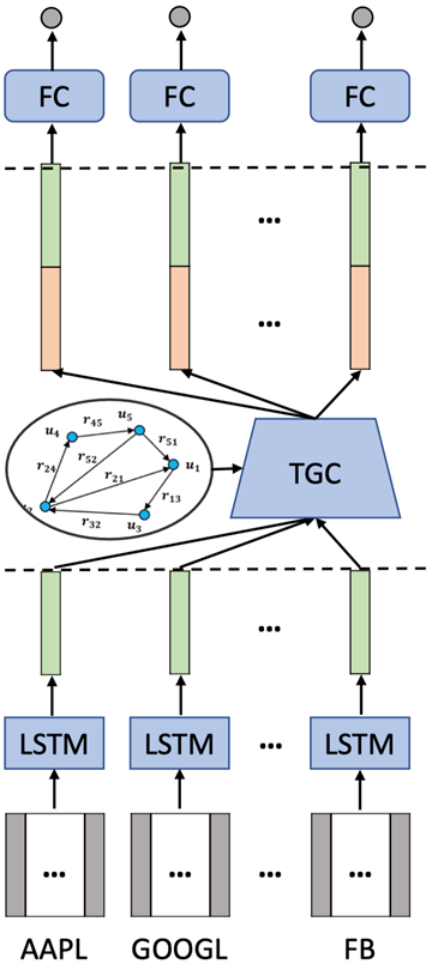

图 4: 图卷积神经网络模型

时序嵌入层：考虑到股票市场的时序动态特征，自然想到使用序列模型预测未来价格趋势。因此，本模型的第一部分是采用经典递归神经网络模型——长短期记忆模型 (LSTM) 来捕捉股票价格的时序特征。相比于其他序列模型，LSTM 模型的一大优势是拥有长期记忆性，这与很多因子对股价有长期影响效应的特征相符合。LSTM 模型基本结构示意图如图5所示。

我们先将包含股票历史价格的时序数据 X_{tj}^{i} 输入网络，t 表示时间截面，j 表示与价量数据有关的特征，i 为股票序号，上 X_{tj}^{i} 表示第 i 个输入在 t 时间的特征矩阵。该过程对应图4的下侧部分。最后我们将 LSTM 的模型的最终隐藏层状态 (h_t^i) 作为时序嵌入层的输出，为了匹配之后图网络的输入。我们有

$$\mathrm{E}^{\mathrm{t}} = \mathrm{LSTM}\left(\mathcal{X}^{t}\right)$$

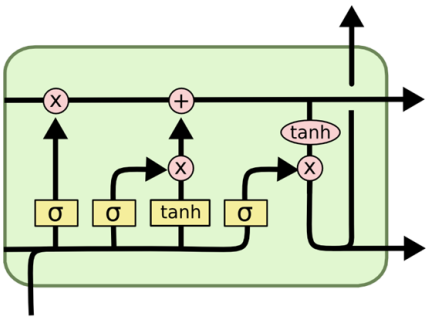

图 5: LSTM 单元结构示意图

其中 $E^t = [e_1^t, \cdots, e_N^t]^T \in R^{N \times U}$ 代表了所有股票的一个序列嵌入 (sequential embedding)，U 是嵌入特征维度 (即 hidden units 的数量)。

关系嵌入层：在此阶段，我们使用时态图卷积神经网络 (Temporal Graph Convolutoin Network, TGC) 以求最好地学习到股票相关性的信息。图神经网络天然地接收图和关系形式的输入，并捕捉到节点与节点之间的关联关系，但是传统的图神经网络，例如图卷积网络 (GCN) 只能捕获到截面数据的关联关系，而忽视了价量数据中丰富的时序信息，因而我们引入 TGC 网络，将 LSTM 网络的时序嵌入通过供应链关系信息进行校准，从而充分考虑到动量和股票上下游关系，得到关系型嵌入的输出。

具体而言，我们将供应链关系数据表示成 $A \in R^{N \times N \times K}$ 的张量形式，前两个维度 N 均为总股票数，第三个维度 K 表示供应链关系总数。例如,$A[i][j][k]$ 表示在第 k 种供应链关系下，股票 i,j 之间的权值大小。该张量作为关系嵌入和 LSTM 输出的时序嵌入共同输入 TGC 网络接受联合训练。该过程对应图4的中间部分。

时态图卷积网络的一大优势是它可以接收和生成随时间动态变化的嵌入。因为公司之间的供求关系不是恒定不变的，所以张量 A 的各元素值是随时间变化的 (相当于有第四个维度 t，不过由于供应链关系在实现上是 3 个月变化一次，我们直接将其看作不同的矩阵)。本文所用代码可以识别输入 LSTM 网络的时序数据对应的时间段，再自动挑选与之匹配的供应链矩阵嵌入 A^t，将其与 LSTM 输出的顺序嵌入一起输入 TGC 网络，真正实现了动态提取股票间的关联关系并融入到对应时段的价量数据中。这也是本文的一个**创新点**所在。

预测层：时态图卷积网络亦会输出一个嵌入 $\overline{E^t} \in R^{N \times U}$，我们再将其输入一个全连接层，即得到最终的输出序列 $F^t \in R^{N \times 1}$ 作为股票排序的打分。基于此得分，我们可以进行选股等后续操作。

2.4 损失函数和评价指标

由于我们的最终目标是得到可供选股的股票得分并进行排名，而并不是真的追求准确预测收益率大小，因此我们在传统均方误差损失（MSE）函数的基础上加入了最大边际损失。根

据 [ZCSZ07]，我们定义整个模型的训练损失函数如下

$$l\left(\hat{\mathbf{r}}^{t+1}, \mathbf{r}^{t+1}\right)=\left\|\hat{\mathbf{r}}^{t+1}-\mathbf{r}^{t+1}\right\|^{2}+\alpha \sum_{i=0}^{N} \sum_{j=0}^{N} \max \left(0,-\left(\hat{r}_{i}^{t+1}-\hat{r}_{j}^{t+1}\right)\left(r_{i}^{t+1}-r_{j}^{t+1}\right)\right)$$

$\hat{r}, r, r$ 分别表示预测值和真值。损失函数中的第一项为均方误差损失，这个传统的回归损失项用于惩罚较大的预测偏差，使得模型的输出结果更加接近真实值的同时，偏差较小。损失函数中的第二项为最大边际损失，最大化了模型对正样本和负样本之间的“间隔”，这有助于增强模型的泛化能力。具体的，对于真值和预测值中，对应序号相同的一对股票，若它们的大小关系相反，则输出 1，若相同则输出 0，用于惩罚不同的预测值和真值对应的相对排序。因此，该项将使得预测值和真值拥有相同的相对排序关系，使用超参数 α 调整最大边际损失的权重，在我们的实验中，设置 $\alpha=0.01$ 的效果较好。

最终模型效果的评判将基于三大指标和多分组回测结果。分别是 Mean Square Error (MSE), Mean Reciprocal Rank (MRR) 和 the cumulative investment return ratio (IRR)。MSE 是经典的平均平方误差, MRR (Mean Reciprocal Rank) 是一个常用的排名性能评估指标。对于 MRR，数值越大越好，取值范围在 0 到 1 之间。MRR 衡量了模型在排序股票时选择的平均倒数排名。较大的 MRR 值表示模型更准确地将高回报的股票排在前面。IRR (cumulative investment return ratio) 则是下一个截面的所选股票收益率总和，直接反映了股票投资的效果。对于 IRR，数值越大越好。而最终的多分组回测结果能最直观地展示选股效果，多空头收益和其夏普比率。

3 数据描述

我们数据的来源是数库 (ChinaScope) 的产业链和供应链数据，我们采用的是供应链数据。该数据中有多种关联关系，而我们主要用的是以下几种：

1. 本公司与本公司的客户与应收、预收账款方
2. 本公司与本公司的关联交易-销售方
3. 本公司与本公司的关联交易-应收账款方
4. 母/子公司关联

以“本公司与本公司的客户与应收、预收账款方”为例，展示部分数据字段如图6所示。

一、对于一级关联，这三个表可以写成矩阵

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

表名：[equity_supplier]- 供应链供应商表

本表记录了以下内容：

供应商：公司主要供应商的情况，包含公司向前 5 名供应商采购额，及其占年度采购总额的比例；

应付账款：公司应付账款情况，包含公司账龄超过 1 年的重要应付账款的应付对象、期末余额；

预付款项：公司预付对象集中度情况，包含公司期末余额前五名的预付款项的期末余额及占预付款项期末余额合计数的比例。

字段	说明	格式	描述
operation	操作状态	VARCHAR(2)	（A：新增/U：更新/D：删除）
id	主键 id	VARCHAR(50)	记录 ID，唯一标识
secu	证券代码	VARCHAR(20)	数库自定义股票标准代码 深市股票以"_SZ_EQ"结尾 沪市股票以"_SH_EQ"结尾 关联[base_stock]"code"
ticker	交易代码	VARCHAR(20)	市场交易代码
rpt	报告日期	VARCHAR(20)	报告截止日（格式：YYYY-MM-DD）
supplier_orig	原始披露名称	VARCHAR(2000)	上市公司披露供应商名称，有时披露公司简称
supplier_id	数库代码	VARCHAR(40)	供应商类型为自然人，关联 [new_base_people]"id"； 供应商类型为机构，关联[std_org]"csfid"；
supplier_cat	供应商类型	INT	0：无效 1：自然人 2：机构
cy_sch	货币中文	VARCHAR(20)	货币中文名称
cy_en	货币英文	VARCHAR(20)	货币英文名称
unit_sch	单位中文	VARCHAR(20)	单位中文名称
unit_en	单位英文	VARCHAR(20)	单位英文名称
amount	金额	DECIMAL(19,2)	Typ: 供应商，采购额 Typ: 应付账款，期末余额 Typ: 预付款项，期末余额

图 6: 供应链关系数据字段示意

其中矩阵的行指标 i 为 A 股上市公司 (股票)，列指标 j 为供应链下游的 A 股上市公司 (股票)。a_{ij} 表示第 j 个下游的股票第 i 个股票的影响。

综上，我们就能得到 i，j 指标都是股票的三个一级转移矩阵，记为

$$\{\hat{T}_{ij}^{n}\}_{n=1}^{3}$$

矩阵元：取两只股票对应的公司间资金流占总供应/销售额的比重；

二、对于二级关联，基于一级关联关系，有以下几种：

1. 本公司与本公司的客户的客户

2. 本公司与本公司的关联交易-销售方的关联交易-销售方

3. 本公司与本公司的关联交易-应收账款方的关联交易-应收账款方

我们可以用 $A * A$ 来获得，下游的下游对本层级的影响。得到 $B = A * A$。

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

b_{ij} 表示第 j 个下游的股票对 (我们关心的) 第 i 个股票的影响。
这就是二级转移矩阵，记为

$$\{\tilde{T}_{ij}^{n}\}_{n=1}^{3}$$

矩阵元：取两只股票对应的公司间资金流占总供应/销售额的比重；

值得注意的是：在得到二阶转移矩阵时，矩阵的对角元可能不为 0，这代表本层级的公司的下游的下游关联到了自己，这也许也代表着什么信息，但是这个信息和其他的关联信息不是一类的，使得转移矩阵在右乘动量-时间矩阵 (M_{jt}，表示 t 时第 j 支股票的动量) 时会将股票的动量传递给自身。我们实际测试发现，不处理对角元得到的 IC 曲线较差，所以我们直接将二阶转移矩阵的对角元赋为 0。

三、对于母/子公司关联，与前面计算一级转移矩阵方法一样，得到 A 股股票的母/子公司的转移矩阵

$$P_{ij}, S_{ij},$$

表示第 j 个母/子公司（股票）对第 i 个本层级公司（股票）的影响。

矩阵元：P_{ij} 取母公司对本公司的持股比例；S_{ij} 取本公司对子公司的总持股比例.

最终得到综合关联矩阵为

$$T = \sum_{n=1}^{3} \hat{T}^{n} + \sum_{n=1}^{3} \tilde{T}^{n} + P + S$$

相加的时候，行列指标都对齐。并将 T 的每一行归一化。之后就利用 T 来做表征股票关系的转移矩阵。

四、总体统计特征：在去除了异常值后，我们得到的关系矩阵维数为 1206×1206，之后的研究过程皆基于这 1206 支 A 股股票。

整个矩阵的值的分布如图7所示，可以看到它是个稀疏矩阵 (受显示限制看不出数值起伏)。矩阵的 cell value(除去 0 之后) 的分布如图8所示，可以看到矩阵的 cell value 分布主要在 0.0 0.2 间，有个别的离群值。

注：

- 我们每 60 个交易日更新一次转移矩阵。
- 矩阵相乘前先将 T_{ij} 的每一行归一化

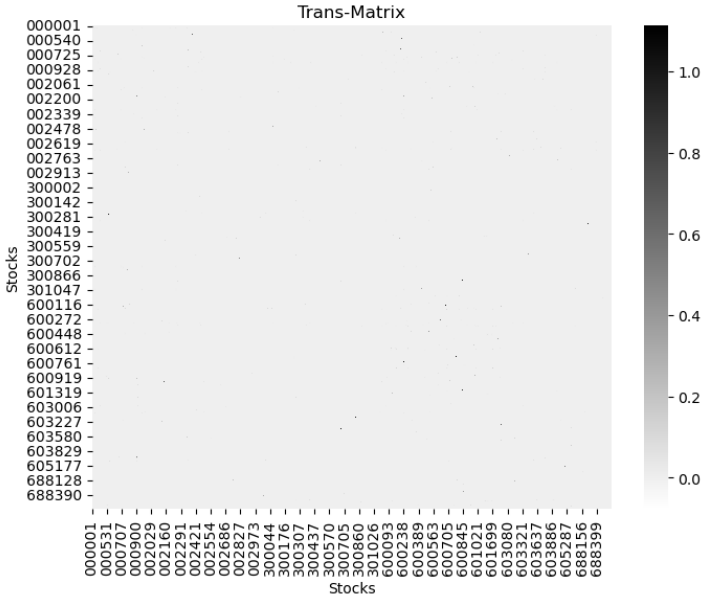

图 7: 供应链矩阵数值分布热力图

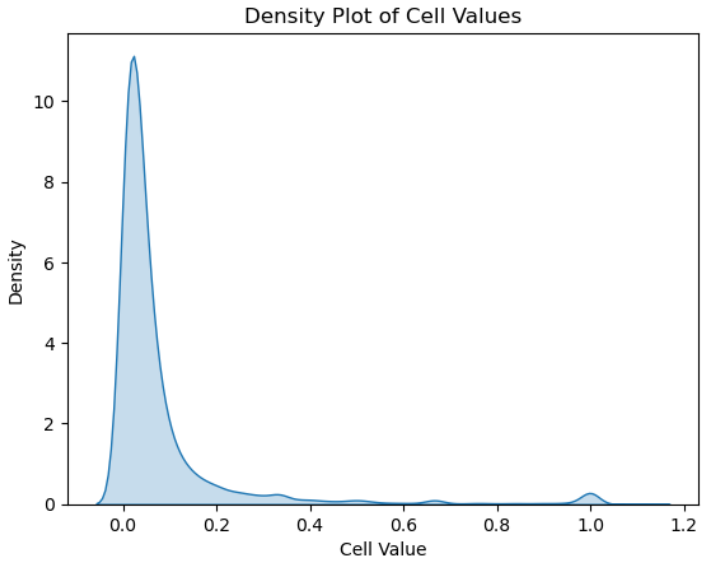

图 8: 供应链矩阵数值分布图

4 实证分析

4.1 供应链数据的应用

4.1.1 简单动量因子传导

根据将上一部分得到的供应链矩阵 T，依据公式

$$F_{it} = T_{ij} M_{jt}$$

我们采用 15 日动量作为传导因子 (M_{jt})，将供应链矩阵左乘于其上，得到新因子 F_{it}。再将新因子对原始动量因子做回归，以提取出纯由供应链关系带来的信息增量。分别按未来 1 日，5 日，20 日收益率在 2017/01/01 至 2021/12/31 这一时间段对回归后的因子 $\widetilde{F}$ 进行回测，结果如下：

	1D	5D	20D
平均收益率 (%)	0.0097	0.0417	0.1081
收益率序列 t 值	7.03	7.94	7.96

表 1: 因子收益率统计

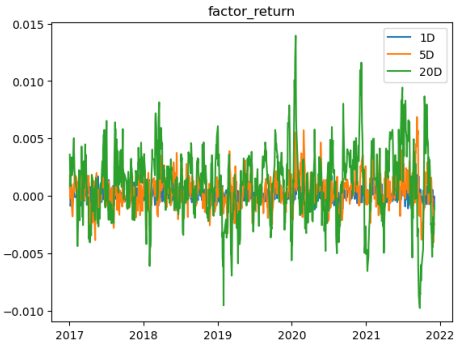

图 9: 每日因子收益率

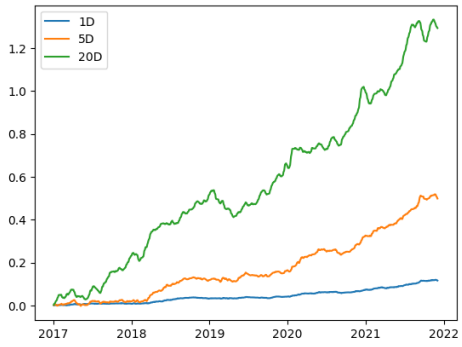

图 10: 累积因子收益率

表 2: IC 值统计

	1D	5D	20D
平均 IC(%)	0.445	0.744	1.027
ICIR	0.13	0.18	0.25

以上结果表明去除动量效应后的供应链动量因子有一定的选股效果。其因子收益率显著不为 0，累积 IC 曲线基本单调，而五分组回测曲线皆有保序性。这说明供应链矩阵自身有一定的可供预测未来股价的作用，在供求矩阵中蕴含着市场并非完全有效的信息。

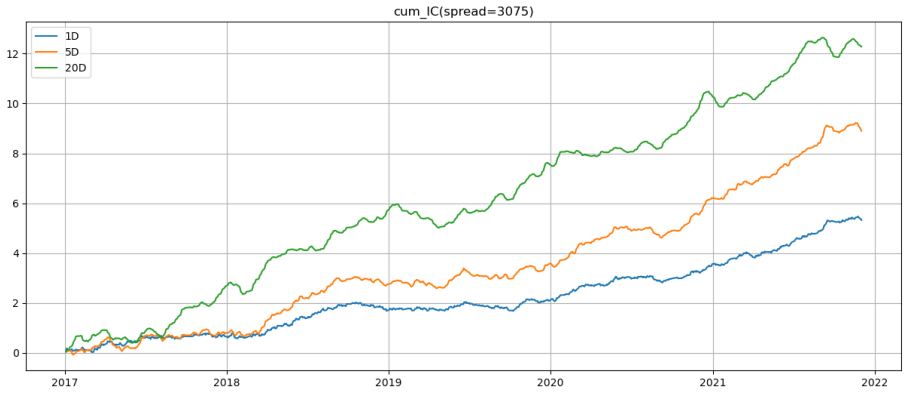

图 11: 累积 IC 值曲线

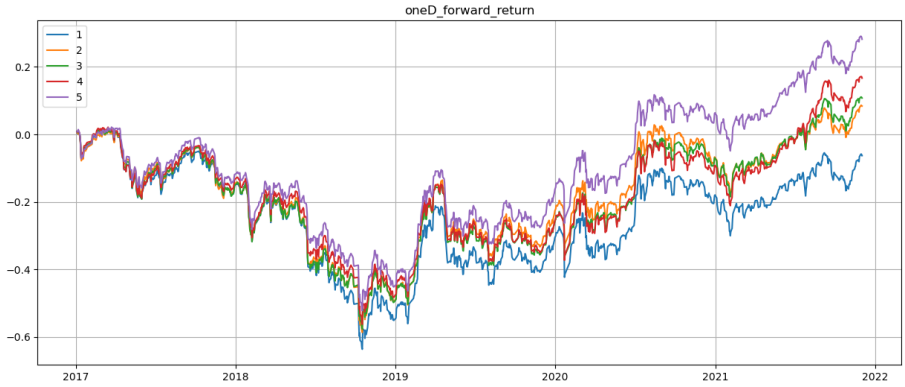

图 12: 1 日收益率回测曲线

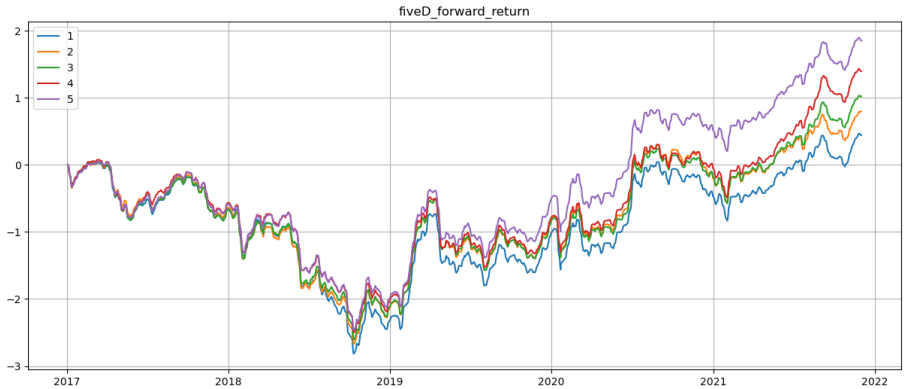

图 13: 5 日收益率回测曲线

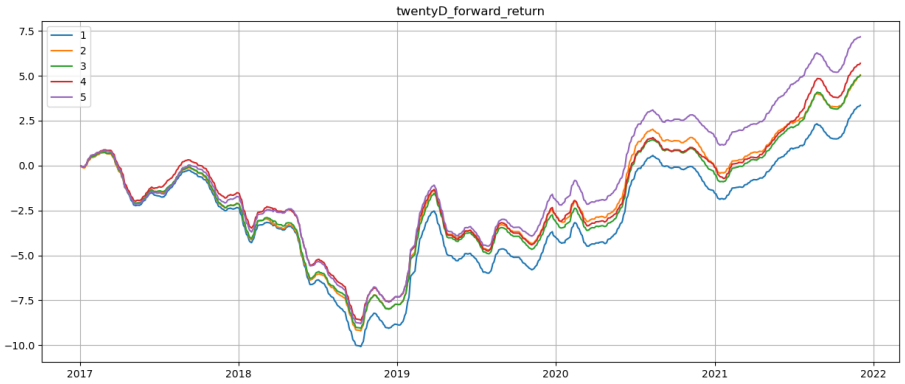

图 14: 20 日收益率回测曲线

考虑到我们仅将八种不同供应链关系矩阵直接相加，这种简单直接的做法有一定的改进空间。下面我们从两个方面优化供应链矩阵：各个转移矩阵相加合成综合转移矩阵时各个矩阵前的系数可以调整、每个矩阵的矩阵元取值可以修改。

4.1.2 供应链矩阵改进

我们之前得到的综合关联矩阵为

$$T = \sum_{n=1}^{3} \hat{T}^n + \sum_{n=1}^{3} \tilde{T}^n + P + S$$

我们尝试找到一个较为合适的权重代替直接相加。由于不同的供应链数据可能对因子有着不同程度，甚至相反方向的信息加成，而直接相加会抹灭这些细节信息。因此我们考虑单独测试每个矩阵的 IC 值，采用以每个矩阵的 IC 为权的合成方法，即

$$T = \sum_{i=1}^{8} IC_i \times T_i$$

其中 IC_i 表示用 T_i 单独作为供应链矩阵，乘以因子算出的新因子的回测 IC。IC_i 值如下 (以 % 为单位)：

表 3: 各矩阵 IC 值

IC_1	IC_2	IC_3	IC_4	IC_5	IC_6	$\|IC_7$	IC_8
0.4790	0.7272	0.5349	0.5037	0.4666	0.3999	0.1710	-0.0176

此外，我们还对如何选取供应链矩阵元做了改进。由于这部分工作主要在尝试各种有关供应链基本面数据的字段，过程较为繁琐枯燥，其实现细节不在报告中呈现。请见存于北大网盘的代码文件。

再次进行回测，得到结果如下页所示

结果表明对供应链矩阵的生成细节优化以后，各回测指标 (平均收益率、IC 值等) 有了全面提升。我们将用本部分得到的供应链矩阵进行后续试验。

表 4: 因子收益率统计

	1D	5D	20D
平均收益率 (%)	0.0104	0.0496	0.1359
收益率序列 t 值	5.51	7.04	11.55

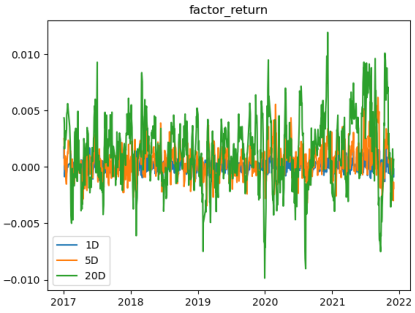

图 15: 每日因子收益率

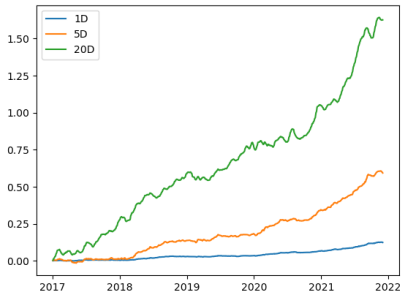

图 16: 累积因子收益率

表 5: IC 值统计

	1D	5D	20D
平均 IC(%)	0.458	0.854	1.243
ICIR	0.23	0.21	0.32

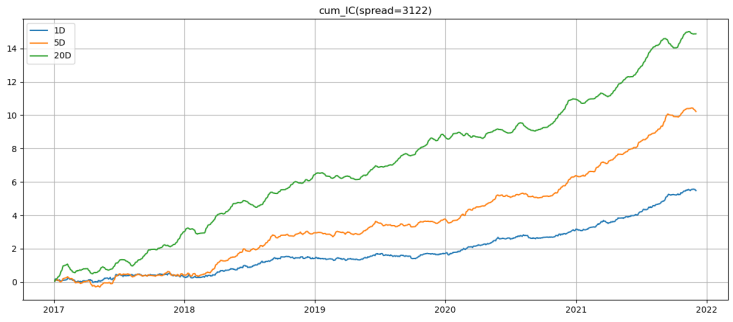

图 17: 累积 IC 值曲线

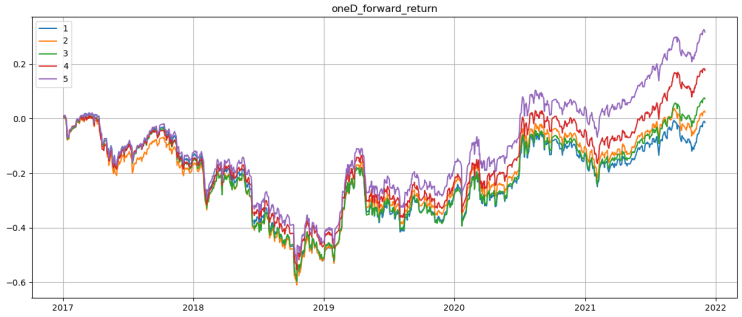

图 18: 1 日收益率回测曲线

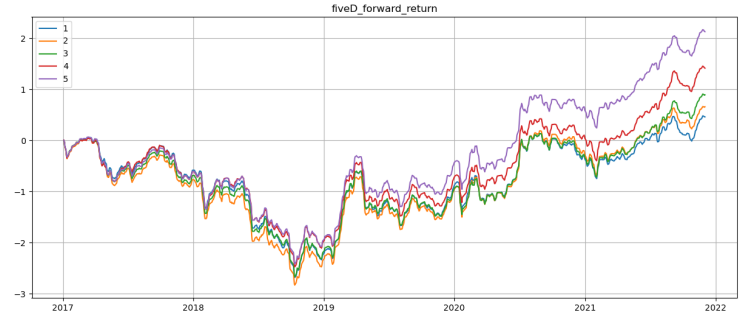

图 19: 5 日收益率回测曲线

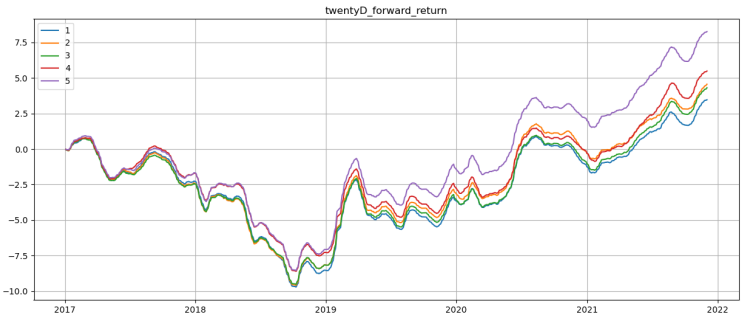

图 20: 20 日收益率回测曲线

4.1.3 供应链因子扩充

我们接下去考虑修改 $F_{it} = T_{ij}M_{jt}$ 右边的 M_{jt}，即经由供应链传导的因子。

我们测试了市值类因子、波动率类因子以及从 worldquant101 公开因子中找到的由遗传算法生成的量价因子，发现经供应量传导后的效果并不理想。有些因子本身效果很好，IC 和 ICIR 都很高，但是经由转移矩阵传导后再回归掉因子本身后得到的因子（来源于矩阵本身供应链信息）效果会很差。这一方面佐证了来源于供应链信息的因子与原本的量价因子相关性很小，但另一方面也给我们寻找其他有供应链溢出效应的因子带来了困难。

为此，我们重新考虑动量因子，但是这次我们对动量因子的时间以及分布都做了修改。比如我们采用 12 个月的动量，并将绝对值大于 0.7 的动量值都赋值为 +-1，这样做强调了大涨或者大跌的影响。除此之外，我们还对分布做了拉伸、消除 0 附近值等 7 种操作（具体测试结果见代码文件夹中的测试 4 文件夹）。图22-图24为 3 个例子。

我们测试了 3M、6M、9M、12M、2Y、3Y 时间上的动量，并挑选了每个时间上 IC 均值最高的分布变形（这些动量的效果有些会与原本的 15D 动量相差不大）。我们将这些动量和原来的 15D 动量，一共 7 个因子，计算他们的相关系数矩阵，如图21所示。可以看到相关性并不是很大，可以放心同时使用这些因子而不担心共线性的影响。

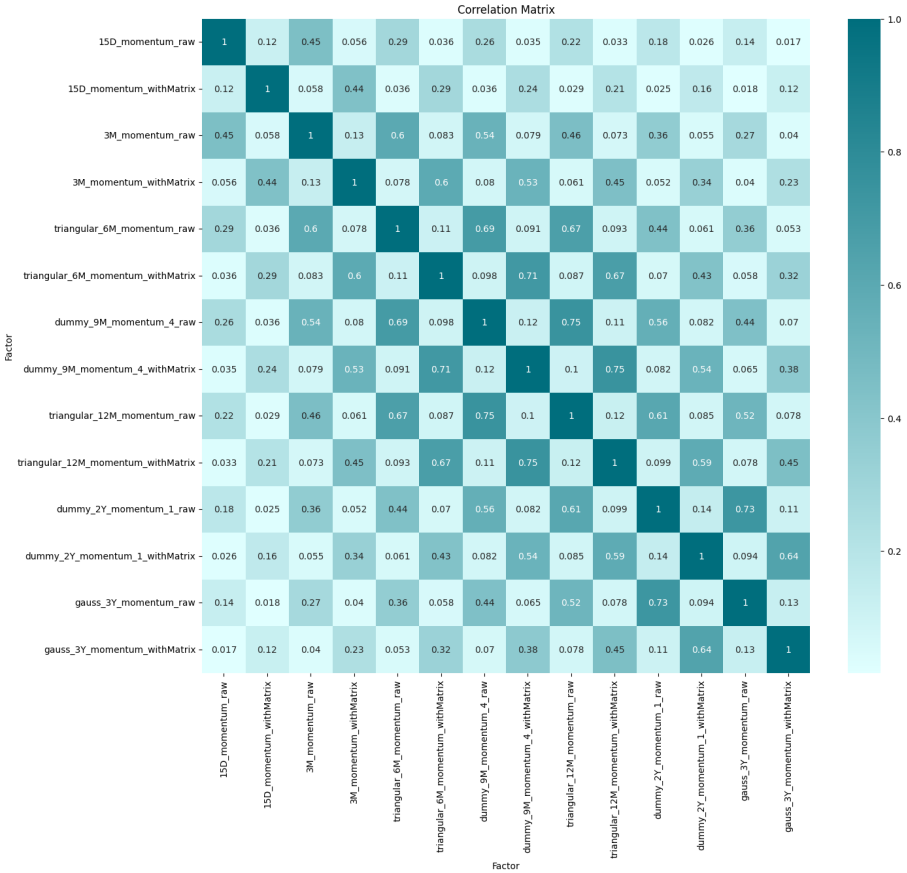

图 21: 7 个因子相关性热力图

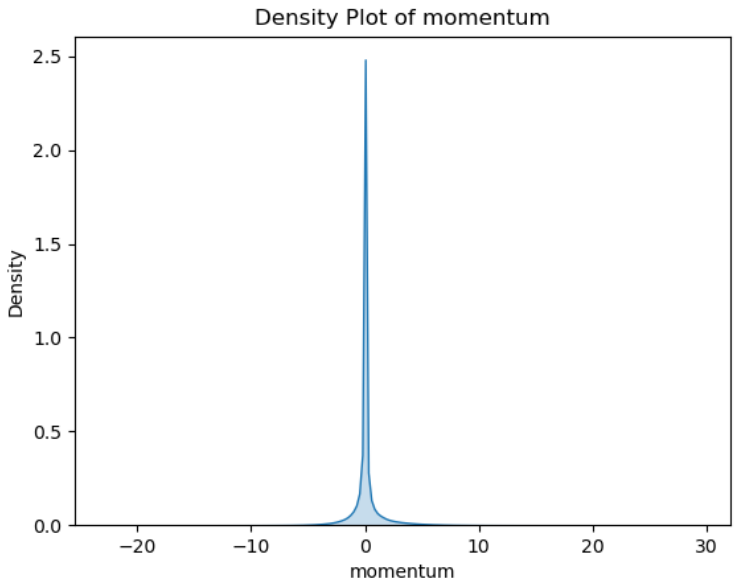

图 22: 对分布进行了拉伸的 3M 动量

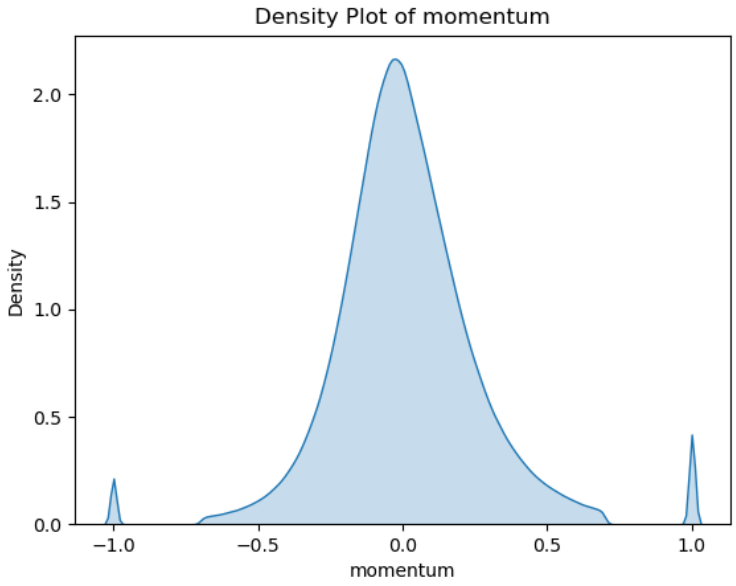

图 23: 对分布极端值进行加大的动量

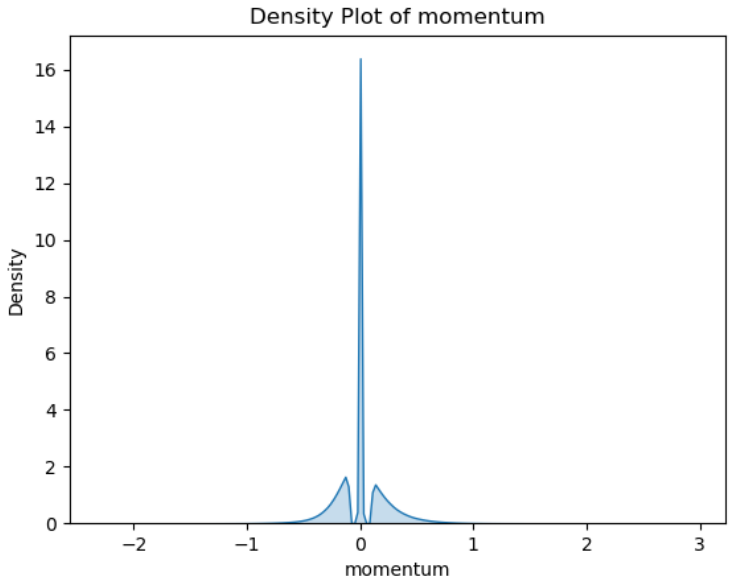

图 24: 对分布 0 附近消除的动量

限于篇幅长度，这些因子具体的回测绩效将不再展示，请详见代码文件夹。这些因子在回测时均能展现有效性和一定的选股能力。

至此，我们已经从供应链数据中提取并确定了供应链关系的转移矩阵，并测试了初步的供应链传导效果，找到了效果较好的 7 个因子。下一部分，我们将采用深度学习的方法，在 LSTM 输入部分加入这 7 个因子，利用模型更充分地学习得到因子组合后的效益。

4.2 深度学习模型的建立与训练

在论文《Temporal Relational Ranking for Stock Prediction》[FHW+19] 中，Relational Stock Ranking（RSR）被证明是一种有效的股票预测方法。RSR 将股票预测视为一个排名任务，并通过在纳斯达克和纽约证券交易所市场上的优异表现，超越了时序网络如 SFM，排名网络如 Rank_LSTM，以及图神经网络如 GBR 和 GCN。此外，实验还展示了时序图卷积（TGC）方法在提高 Rank_LSTM 预测性能方面的有效性。在这个基础上，我们设计了三个主要的实验方向：

- 由于 RSR 在原论文中仅在美股市场上进行了验证，我们想要进一步探究 RSR 在中国 A 股市场上是否同样有效。
- 我们在模型中引入了一种新的关系类型——供应链关系，想要验证这种新关系能否在 Rank_LSTM 的基础上进一步提升模型性能。
- 我们还将研究基于供应链图关系的 Rank_LSTM 网络在不同回测方案下的表现。

4.2.1 评估指标

为了全面和精确地衡量模型的性能，我们选择了三个主要的评估指标：均方误差（MSE）、平均互换等级（MRR）和累积投资收益率（IRR）。MSE 是回归任务的常用评估指标，如股票价格预测，我们将在测试期间的每个交易日计算所有股票的 MSE。MRR 则是评估排名性能的指标，我们将计算所选股票在测试期间每天的平均倒数排名。IRR 直接反映了股票投资的效果，它的计算方式是累加所选股票在每个测试日的收益率。为了消除不同初始化可能引起的波动，我们将对每种方法进行五次测试，并报告平均性能。

4.2.2 模型方法

本次实验主要采用三种深度学习方法：

- GCN：图卷积网络（GCN）作为一种前沿的基于图的学习方法，我们使用 GCN 层代替 RSR 中的 TGC 层。股票关系图将送入 GCN 进行处理。
- Rank_LSTM：通过去除 RSR 模型中的关系嵌入层，我们获得了 Rank_LSTM 模型，这是一种排名损失的 LSTM 模型，但不考虑股票间的关联关系。
- RSR：该方法为在 TGC 中隐式地对关联关系进行建模的 RSR 模型，根据原论文，此种隐式方法的效果优于显式方法。

4.2.3 参数设置

所有模型均通过 TensorFlow 进行实现。虽然原论文中的代码基于 TensorFlow 1.x 版本，但我们对其进行了修改，使其能够兼容 TensorFlow 2.x，并已在 GitHub 上公开了我们的代码：代码链接。我们使用 Adam 优化器以 0.001 的学习率训练 Rank_LSTM 和 RSR 模型，并通过网格搜索选择最佳超参数。我们在 {2, 4, 8, 16} 和 {16, 32, 64, 128} 范围内分别调整了 LSTM 的顺序输入长度 S 和隐藏单元数量 U，并通过 {0.1, 1, 10} 调整了损失函数中的 α，以平衡点对点项。

4.2.4 数据筛选

考虑到我们的供应链关系矩阵始于 2017 年，因此我们选择了从 2017 年 1 月 3 日至 2021 年 12 月 31 日的 1248 个交易日内的 1206 支股票进行实验。我们按照以下原则从 A 股市场中筛选股票：

1）在选定的时间段内，股票在 98% 的交易日有交易活动；

2）自 2017 年 1 月 3 日起，股票价格从未低于 1 元人民币。最后，我们收集了两种类型的股票数据，包括历史交易数据和时序供应链关系矩阵。

4.2.5 实验探究

表 6: 验证集结果

	MSE	MRR	IRR
VALID			
GCN	9.29×10^{-4}	3.61×10^{-2}	0.64
Rank_LSTM	$\mathbf{8.27 \times 10^{-4}}$	2.25×10^{-2}	0.86
RSR	8.29×10^{-4}	$\mathbf{4.38 \times 10^{-2}}$	**2.86**

表 7: 测试集结果

	MSE	MRR	IRR
TEST			
GCN	9.29×10^{-4}	3.61×10^{-2}	0.98
Rank_LSTM	8.29×10^{-4}	4.51×10^{-2}	3.65
RSR	$\mathbf{7.98 \times 10^{-4}}$	$\mathbf{4.72 \times 10^{-2}}$	**4.87**

我们的实验结果表明，引入股票排名解决方案的 Relation_Rank_LSTM 模型相较于传统的 GCN 模型在 A 股市场的投资回报率（IRR）上取得了显著的提升。这证实了股票排名解决方案在 A 股市场上的优势，并回答了我们之前提出的问题：股票排名解决方案是否同样适用于 A 股市场。结果显示，答案是肯定的。

同时，我们还发现，通过在模型中引入时序图卷积（TGC）组件，Rank_LSTM 模型的预测性能进一步提升，这进一步印证了 TGC 组件在时间序列分析中的有效性。

更为重要的是，我们的实验还发现，引入供应链关系矩阵的 RSR 方法在验证集和测试集的平均倒数排名（MRR）和投资回报率（IRR）指标上，均显著优于 GCN 和 Rank_LSTM。在我们更关注的 MRR 指标上，RSR 模型展现了明显的优势。尽管对于平均平方误差（MSE）这一指标，Rank_LSTM 在验证集上表现得稍优于 RSR，但总体而言，RSR 模型在 A 股市场的表现依然十分出色。

更为值得注意的是，我们所提出的供应链关系矩阵显著提升了 RSR 模型的预测效果，这一结果再次验证了供应链信息在股票预测中的重要性，为未来的股票预测研究提供了新的思路和方向。

4.2.6 分组回测

我们使用测试集表现最好的 RSR 模型和相应超参数所输出的因子值作为最终的因子结果，进行详细的 A 股市场五分组选股回测，结果如图25至27所示。

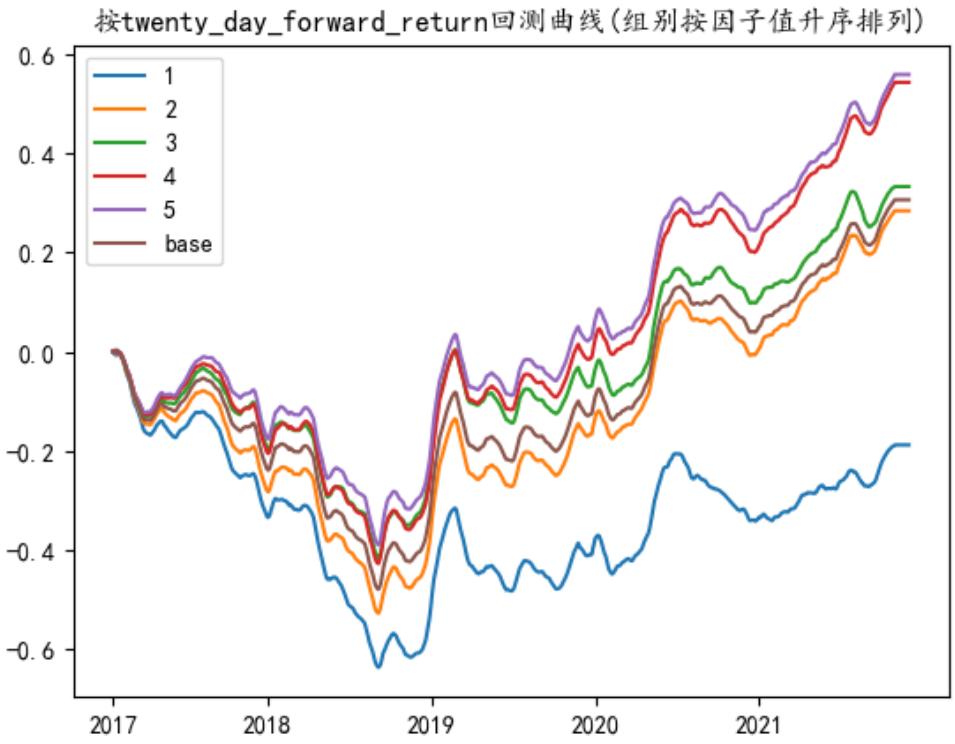

图 25: RSR 模型供应链因子 20 日收益率率五分组回测曲线

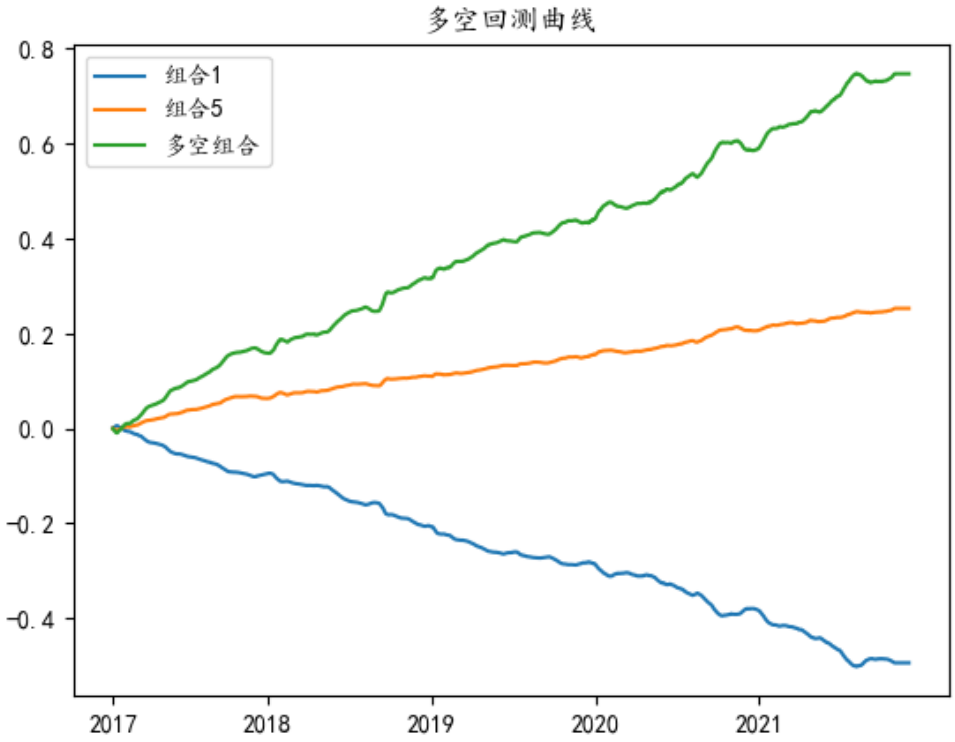

图 26: RSR 模型供应链因子 20 日收益率率五分组多空对冲曲线

回测结果								
	1	2	3	4	5	组合1	组合5	多空组合
年化收益率%	-3.993423	6.052796	7.092269	11.563108	11.899136	-10.516672	5.375887	15.892559
年化波动率%	4.699891	4.869945	4.922751	5.029492	4.992910	1.054968	0.670805	1.578719
夏普比率	-1.275226	0.832206	1.034436	1.901406	1.982639	-11.864508	5.032589	8.799895
最大回撤%	-63.924899	-52.740596	-41.389121	-42.910727	-38.924176	-50.662536	-0.853143	-2.101604
胜率%	45.562130	51.310228	53.085376	54.353339	54.860524	23.245985	70.414201	74.978867

图 27: RSR 模型供应链因子 20 日收益率率五分组回测绩效

多分组回测结果同样具有保序性,最高的多头和最低的空头年化收益分别为 5.38%, −10.51%, 多空对冲组合有 15.89% 的年化收益。多头收益并不算高，但其波动率相对较低，因此拥有整体很高的夏普比率和胜率，最大回撤也比较好看。在实验过程中，确实发现供应链因子含有的信息增益 alpha 不算太强，但信号较为稳定的性质。至此，我们通过模型构造出了独立于股票量价因子之外的，充分利用供应链关系数据的单 alpha 因子———可命名为供应链传导 RSR 因子。

5 结论

本文利用了 2017 2021 年中国 A 股 1206 家上市公司的股票价格基本面数据, 以及相同时间段内表征公司供求关系的供应链基本面数据, 自主构造了供应链关系矩阵，基于不同的传统价量因子生成了供应链传导因子并检验了其有效性。之后利用由 LSTM 时序循环神经网络和 TGC 时态图卷积神经网络组成的 RSR 模型，综合学习了股票价格基本面数据和供应链关系数据，训练出了最终的深度学习因子，并进行五分组等权回测，获得了多空对冲年化收益率超过 15%，夏普率超过 8，最大回撤 −2% 的稳健良好的回测效果。

值得指出的是，构建供应链矩阵的方式可能有很多种，亦有许多实现细节尚未探究。此外，多分组回测并未考虑交易成本与手续费，在构建组合时也只是简单等权组合，未考虑行业等风格因素的影响。然而这些初步的研究已可以证明供应链数据含有未被发掘的有效信息和市场的弱有效性，本文获得的单因子也显然拥有被有效运用到更大范围的多因子模型中的潜力。

参考文献

[CF08] LAUREN COHEN and ANDREA FRAZZINI. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011, 2008.

[FHW+19] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Trans. Inf. Syst.*, 37(2), mar 2019.

[MG99] Tobias J. Moskowitz and Mark Grinblatt. Do industries explain momentum? *Journal of Finance*, 54(4):1249–1290, 1999.

[PST20] Christopher A Parsons, Riccardo Sabbatucci, and Sheridan Titman. Geographic Lead-Lag Effects. *The Review of Financial Studies*, 33(10):4721–4770, 01 2020.

[PTX15] Li Pan, Ya Tang, and Jianguo Xu. Speculative Trading and Stock Returns*. *Review of Finance*, 20(5):1835–1865, 12 2015.

[Ras01] Michael S. Rashes. Massively confused investors making conspicuously ignorant choices (mci–mcic). *The Journal of Finance*, 56(5):1911–1927, 2001.

[ZCSZ07] Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[张 22] 段丙蕾, 汪荣飞 , 张然. 南橘北枳：a 股市场的经济关联与股票回报. 金融研究, No.500(171-188), 2022.

[谭 22] 谭晶桦. 面向公司关联的多因子实证资产定价研究. 2022.