个股研究报告—专题报告

金融工程

数量化投资

学术文献研究系列第 16 期

2021 年 11 月 25 日

金融工程报告

相关研究报告:
《港股投资周报：南向资金短期流出，港股精选组合今年以来超恒生指数 19.54%》 ——2021-11-20
《数量化投资周报：多因子选股周报—估值类因子表现出色，中证 500 指数增强组合本周超额 1.13%》 ——2021-11-20
《主动量化策略周报：持续反弹，超预期精选组合今年以来满仓收益 48.72%》 ——2021-11-20
《ETF 周报：大金融板块躁动隐现，首批增强 ETF 开始发行》 ——2021-11-22
《金融工程专题研究：新规下的沪深核心指数成分股调整预测》 ——2021-11-21

证券分析师：张欣慰
电话：021-60933159
E-MAIL：zhangxinwei1@guosen.com.cn
证券投资咨询执业资格证书编码：S0980520060001
联系人：杨北锋
电话：021-60875136
E-MAIL：yangbeifeng@guosen.com.cn

构造多空股票投资组合：一种新的排序学习算法

● 研究背景

学习排序算法是一类有监督的机器学习算法，基于排序的机器学习算法在文本归纳和机器翻译等领域也表现出了强大的能力。然而在因子投资领域这些排序算法还没有得到人们的足够重视。另外对于 IR 指标，在多空策略中人们希望排名靠前和靠后的指标都是准确的。

● 研究现状

将排名框架引入到了多因子策略大多关注于添加可替换因子或构建神经网络。Song 将排名进行颠倒，并对模型进行两次拟合，希望两个模型分别能够预测头部和尾部。但是这种方法可能得出在同一时间买卖同一只股票的结果。这样的结果促使本文提出了一种新的学习排序算法，该算法基于头部和尾部一致的排序列表，并且在排序列表中两者同等重要。

● 损失函数介绍

在多空策略的启发下，考虑到投资组合的收益和损失，本文提出了一种新的替代损失函数 ListFold，该损失函数将队列的头部和尾部视为同等重要，并且具有平移不变性以及和 0-1 损失函数的一致性。

● 实证研究

数据集包含了中国 A 股市场 3712 只股票的 631 周的数据，其中包含 68 个因子，样本日期为从 2006-12-29 到 2019-04-19。模型采用了 4 层全连通神经网络结构，采用 ListFold 与 MLP、ListMLE 和 Song 拟合的 List2MLE 作为损失函数，分别构造两种不同的多空组合进行比较。一种是做多头部 10%的股票，同时做空尾部 10%的股票；另一种是做多头部 10%的股票，做空所有股票的平均值。对相同投资方向的股票分配相同的权重，每周投入 1 美元。

对比结果显示，自 2016 年以来，做空尾部的情况下，ListFold-exp 的收益最高，MLP 的净值仅为 ListFold-exp 净值的 2/3，ListFold-exp 优于 List2MLE。做空平均值的情况下，ListMLE 和 MLP 收益甚微，而 ListFold 仍然能够获得不错的收益。

最后设定年化无风险率为 3%，每笔交易的总交易成本为 30 个基点，所构建的采用 ListFold 损失函数的模型做空尾部 10%股票投资组合表现较好，样本外年收益率 38%，夏普比率为 2。最后进行了敏感性测试，结果证明策略在回测期市场动荡的环境下表现稳定。

● 结论

本文在学习排序方法的基础上提出的新的损失函数 ListFold，可以被看作是研究非顺序敏感损失函数的一个补充工具，它也可以启发统一 pairwise 和 listwise 代理损失函数的框架。此外，本文验证了 IC 比 NDCG 类型的排名指标更适合评估 alpha 策略。

风险提示：本报告基于相关文献，不构成投资建议。

内容目录

图表目录

文献来源

文献来源: XIN ZHANG, LANWU and ZHIXUE CHEN, “Constructing long-short stock portfolio with a new listwise learn-to-rank algorithm (July, 2021)”, Quantitative Finance.

文献亮点:

本文提出了组合因子构建多空组合的新思路，在学习排序方法的基础上提出了一种新的损失函数，目的是通过 listwise 方式选择多空组合对。它具有平移不变和概率可解释的特点。本文构建的模型可以被看作是研究非顺序敏感损失函数的一个补充工具,它也可以启发统一 pairwise 和 listwise 代理损失函数的框架。

本文的实证结果显示出排序预测相对于数值预测的优势，还有构建的损失函数特别是 ListFold-exp 的强大作用。本文还通过经验验证了 IC 比 NDCG 类型的排名指标更适合评估 alpha 策略。

摘要

随着机器学习的快速发展，多因子策略在行业中的应用越加广泛。通常的做法是将多个因子输入到算法中来做横截面收益的预测，并进一步构建多空投资组合。最新的研究不再是预测股票的收益值，而是利用成熟的排序算法来预测股票的未来收益排名。

本文提出了一个损失函数基于排名的全新的 listwise 排序算法,该损失函数旨在同时强调排名队列的头部和尾部。本文的损失函数用于多空策略具有内在的平移不变性，可以看作是对 ListMLE 的直接推广。在不同的转换函数下，二元分类损失或排列级别的 0-1 损失是一致的。作者也给出一个广义的 Plackett-Luce 模型的概率解释。本文测试了 2006-2019 年中国 A 股市场 68 个因子的数据集，所构建的多空股票投资组合表现较好，样本外年收益率 38%，夏普比率为 2。

关键词: Learn-to-rank ListMLE 多空组合 因子策略 机器学习 JEL 分类 G11 C45 C53

1. 介绍

主动组合管理最早可以追溯至 CAPM 模型（Sharpe，1964 和 Lintner，1965）和 Fama-French 三因子模型（Fama 和 French ，1993）的提出。自打诞生以来，关于主动组合管理的价值众说纷纭，褒贬不一。Carhart(1997)代表传统投资理念的巅峰，Carhart 并不符合擅长技术分析或者精通消息的基金组合经理的条件，但是 Cremer 最新发表的文章（Cremer et al.2019）质疑了传统投资理念这个观点，文章认为无论是在择时还是选股上，传统观点都有点落伍。择时主要包括对市场未来走势的预测，预测周期从几毫秒到几个月不等。而选股策略的一个基本的框架是找到各种因子，并将它们组合起来预测下一个时间区间的回报。在 Fama 和 French 的启发下，为了获得这种基于横断面预测的回报，投资者可以采用因子多空策略，即做多因子排名靠前的股票，做空因子排名靠后的股票。预测能力强的因子往往是保密的，并悄悄获取不菲的收益。与此同时，各类文献中也提出了许多相对普通的因子（Tulchinsky 2019，Giglio 等，

2019）。能否从这些相对普通的因子中挖掘出能产生预测能力强的因子，是一个仁者见仁智者见智的问题。本文试图通过一种新颖的学习排序算法来回答这个问题，该方法受到多-空交易行为的启发，旨在直接预测一个横截面上的股票排名列表，而不是传统的股票收益的数值。

学习排序算法是一类有监督的机器学习算法，在信息检索(IR)领域已经被证明是非常成功的。排序算法已成为推荐系统的核心部分，广泛应用于网络搜索、新闻推送、在线购物和广告等领域。基于排序的机器学习算法在文本归纳和机器翻译等领域也表现出了强大的能力。然而，在因子投资领域这些排序算法还没有得到人们的足够重视。尽管已经有许多其他的机器学习算法被用于开发多因子策略（De Prado 2018，Rasekhschaffe 和 Jones 2019），但是并没有从排名的角度来看待这个问题。对于 IR 指标一个明显的差别是，在多头策略中人们只关心排名靠前的指标的准确性，但在多空策略中，人们希望排名靠前和靠后的指标都是准确的。为了弥补这一差距，人们需要一种同时强调顶端和底部的学习排序算法。事实上，一些研究人员经将排名框架引入到了多因子策略中，但他们大多关注于添加可替换因子或构建神经网络（Song et al. 2017,Feng et al. 2019, Fang et al. 2020a）。对于满足多空策略的需求，很少有人关注到这一点。在 Song et al.(2017)中，作者将排名标签颠倒，并对模型进行两次拟合，希望两个模型分别能够预测头部和尾部。尽管参数调优很麻烦，但是这种方法可能从两个模型得到相反的结果。因此，这种方法不能保证一个连贯的排名，而是可能得出在同一时间买卖同一只股票的结果。这样的结果促使我们提出了一种新的学习排序算法，该算法基于一个一致的排序列表，并且在排序列表中头部和尾部同等重要。

本文并不着重考虑股票的绝对收益，而是侧重于预测收益的相对排名。这一点不仅可以用基金经理跑赢相对指数的目标来解释，也可以通过做出准确价值预测的难度来体现。预测的困难主要来自于输入信息的边界模糊和财务数据的低信噪比。比如一夜之间的突发消息可能会严重影响整个市场，一些不可预测的交易行为影响了股票价格。目前尚不清楚是否有专业人士或因子能够可靠地预测个股的回报，但是我们更倾向于预测股票相对于其他股票或因子模型的走势。Song et al.(2017)、Feng et al.(2019)、Zhu et al.(2011)、Wang and Rasheed(2018) 也认同预测收益排名的观点，并且使用排序信息系数(rank information coefficient, IC)来作为评级指标。

本文的研究结果具有两面性。首先，我们为一个多空策略开发了一个新类型的学习排序损失函数。我们可能是第一个提出了一个基于自上而下和自下而上的学习排序算法框架的多空投资策略。特别地，我们设计了一种新的替代损失函数，并讨论了它的理论特点。我们构造的损失函数由于其对称性质而具有平移不变性，在不同的转换函数下，可以推导出它与二元分类损失或排序的 0-1 损失相一致，我们还对模型进行了概率解释。其次，我们进行了详细的实证研究，检验我们的模型在中国 A 股市场的表现：年化收益率为 38%，夏普比率为 2。可以看出，我们的方法优于多层感知(MLP)、ListMLE 和 Song 两次使用的 ListMLE 模型。

本文的其余部分组织如下。第 2 节简要介绍了背景知识，包括多因子策略、学习排序算法框架和 ListMLE 算法。第三部分给出了我们的模型及其理论分析。第四部分对我们的模型和其他模型在中国 A 股市场上的应用进行了实证比较。在第五部分，总结了我们的主要内容和未来研究的一些想法。

2. 研究背景

2.1 多因子策略

因子是多因子策略的核心。各类已经发表的文献构造了数百种潜在的定价因子(Harvey et al. 2016, Hou et al. 2017)，在行业中甚至更多。一类多因子策略是在一个多重测试框架中反复测试这些因子(Feng et al. 2020b)，但另一类策略，是从因子库中生成效果更好的因子，因而更受欢迎。因子组合最传统的方法可能是顺序过滤、打分法和线性回归。计量经济学家和统计学家从 SVM、Adaboost 和图模型等不同模型的角度进一步扩展了线性模型(Liu et al. 2016)。机器学习同样被用来生成因子和构建因子组合。例如，通过文本分析，机器学习能够从新闻中提取情感因子。在构建因子组合方面，这些算法通常将因子视为输入，将收益视为输出，将问题转化为分类或回归问题。

当投资者想要投资估值过高的标的时，多空策略是纯多头投资的一种自然的替代。Jacobs 和 Levy(1993)研究了多空策略的构建方式、原理和实际收益，以及多空策略在实际运作中产生的问题。如今，对冲基金广泛使用多空策略构建投资组合。另一种比较流行的多空因子的投资方式是先把股票分组，比如总共分成 10 组，然后做多排名靠前的组，做空排名靠后的组。由于不需要估计因子载荷，分组法受到了人们的欢迎。

2.2 学习排序算法概述

学习排序算法源于网页搜索，同时包含了很多以前的方法。它的基本框架是将文本（在本文的场景中是股票）及其特征作为输入，并将基于相关性判断的排名列表作为输出。我们使用一个替代损失函数来学习一个评分函数，它根据文本的特征给每个文本分配一个分数，这样只要我们根据它们的分数对文本进行排序，我们就可以很好地预测它们的排名。计分函数不会因文本而不同。根据损失函数的不同，大多数学习排序算法可以分为点排序、成对排序和列表排序。

为了评估预测结果，特别是关心排名是否正确时，NDCG 是最受欢迎的指标，其定义如下:

$$NDCG@K(\pi, l) = \frac{1}{Z_k}\sum_{j=1}^{k} G(l_{\pi^{-1}(j)})\eta(j) \qquad (1)$$

其中π为预测列表，$\pi^{-1}(j)$表示位于π列表j位置的文本。l代表相关性判断，$G(z) = (2^z - 1)$是文本的常用评级函数。$\eta(j)$是一个位置折扣因子(通常设置为$1/log(1+j)$。位置k的截断意味着我们只关心排名前k个位置的准确性。Z_k是设置 NDCG 落在[0,1]范围内的归一化值。因此我们可以用 1 – NDCG 来代表实际的损失。此外，当且仅当两个列表相同时，排列级别 0-1 的损失达到 0。二进制分类损失是首先将排名列表的前 50%标记为 1，其余为-1。在众多排名指标中，GAUC (Song and Meyer 2015)也考虑了排名列表头部和尾部的准确性，但他的局限性在于只考虑了{1，0，−1}标签的特殊情况。

2.3 ListMLE

ListMLE 是一种先进的列表学习排序算法，它以自上而下的方式定义基于 Plackett-Luce 模型的概率分布。其目的是利用一种可能性损失作为替代损失，定义为:

$$\mathcal{L}(f,x,y) = -log\mathbb{p}(y|x;f) = -log\prod_{i=1}^{n}\frac{\psi\left(f\left(x_{y^{-1}(i)}\right)\right)}{\sum_{k=i}^{n}\psi\left(f\left(x_{y^{-1}(k)}\right)\right)} \qquad (2)$$

其中$y^{-1}(i)$表示标记在第i个位置的分量。x为特征向量，n为样本量，f为打分函数。函数$\psi(\cdot)$是将得分映射到$\mathbb{R}^+$的变换函数。变换函数$\psi(\cdot)$通常被认为是线性的、指数的或 s 型的。为简单起见，$\psi_i \coloneqq \psi\left(f\left(x_{y^{-1}(i)}\right)\right)$。

Plackett-Luce 模型的概率解释是 Silverberg(1980)给出的花瓶模型比喻。考虑从一个装满彩色球的花瓶里画球。每种颜色的球数与ψ_i成比例。假设有无限个球，如果需要不合理的比例。在第一阶段，从花瓶中取出一个球 c1，这种选择的概率是$\psi_1/\sum_{i=1}^{n}\psi_i$。在第二阶段，绘制另一个球，如果它与第一个球的颜色相同，然后把它放回去，并继续尝试，直到选择一个新的颜色 c2；第二次选择的概率为$\psi_2/\sum_{i=2}^{n}\psi_i$。继续这些步骤，直到每个颜色的球都被选中。颜色序列的概率如式(2)所示。

3.我们的模型

在本节中，我们提出模型 ListFold。在多空策略的启发下，我们提出了一种新的替代损失函数，将队列的头部和尾部视为同等重要的，因为它们都有助于投资组合的收益和损失。为了不失一般性，我们假设给偶数个股票（记录）来排序。

3.1 ListFold

对于 2n 个记录X_1，…， X_{2n}，观察排序 y 和评分函数 f，我们尝试将一个排列分解为一个有序的分步骤选择过程：第一个多空记录对，第二个多空记录对，直到第 n 个多空记录对。我们可以这样定义概率公式：

$$\mathbb{P}_c(y|X,f) = \prod_{i=1}^{n}\frac{\psi(f_i - f_{2n+1-i})}{\sum_{i\le u\ne v\le 2n+1-i}\psi(f_u - f_v)}, \qquad (3)$$

其中$f_i \coloneqq f(X_{y^{-1}(i)})$表示在第$i$个位置观察到的股票的得分，$\psi$为转换函数，比如 ListMLE。然后将损失函数定义为负对数似然估计：

$$\begin{aligned}\mathcal{L}_c(f,y,X) &= -\log\mathbb{P}_c(y|X,f)\\ &= -\sum_{i=1}^{n}\left(\log\psi(f_i - f_{2n+1-i}) - \log\sum_{i\le u\ne v\le 2n+1-i}\psi(f_u - f_v)\right).\end{aligned} \qquad (4)$$

我们构造的损失函数的考虑类似于 ListMLE，它将置换概率分解为逐步条件概率。不同之处在于，对于每一步，我们的目标不是选出一个股票，而是选出得分差异最大的一对股票，并进一步将它们按正确的成对偏好顺序排列。根据前面$i-1$步骤的正确排序，我们可以将条件概率写成：

$$\begin{aligned}\mathbb{P}_c^i &:= P_i\left(y^{-1}(i,2n+1-i)\middle|X,y^{-1}(1,2n),\ldots,\right.\\ &\left. y^{-1}(i-1,2n-i+2);f\right)\\ &= \frac{\psi(f_i-f_{2n+1-i})+\psi(f_{2n+1-i}-f_i)}{\sum_{i\le u\ne v\le 2n+1-i}\psi(f_u-f_v)}\cdot\\ &\quad\times\frac{\psi(f_i-f_{2n+1-i})}{\psi(f_i-f_{2n+1-i})+\psi(f_{2n+1-i}-f_i)}\\ &= \frac{\psi(f_i-f_{2n+1-i})}{\sum_{i\le u\ne v\le 2n+1-i}\psi(f_u-f_v)},\quad i=1,\ldots,n. \end{aligned}\tag{5}$$

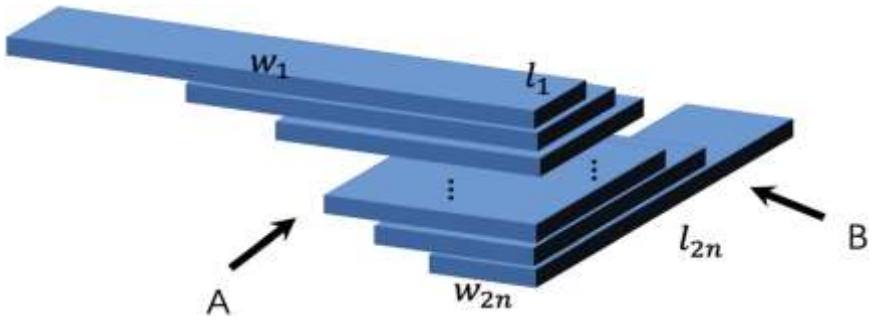
这就很自然地解释了$\mathbb{p}_c$中分母的对称性。而且由于这种对称性，ListFold 天生就是平移不变的，而 ListMLE 只有当ψ是指数时才会是平移不变的。

当ψ为指数时，我们的损失概率解释可由花瓶模型的自然泛化推导出来。考虑一个投掷飞镖的多级实验。有 2n 块木板堆在一起。宽度、长度和高度分别为(w_i, l_i, 1)，约束条件为w_i*l_i=1。在第一阶段，A 和 B 同时各自向板的宽度和长度方向随机投掷飞镖。如果他们的飞镖落在同一块木板上，那么把木板放回去，再投一次，否则标记 A 板为A_1，B 板为B_1。继续进行重复实验，不要把标记的木板放回去，在步骤i中标记木板为A_i和B_i，直到所有的木板都标记好。然后木板序列的概率{A_1，…，A_n，B_1，…，B_n}:

$$\mathbb{P}=\prod_{i=1}^{n}\mathbb{P}_i:=\prod_{i=1}^{n}\frac{w_{A_i}*l_{B_i}}{(\sum_{j=i}^{n}w_{A_j}+w_{B_j})*(\sum_{j=i}^{n}l_{A_j}+l_{B_j})\atop -2(n+1-i)},\tag{6}$$

这是方程(4)中假设ψ指数和f_i=log(w_i)，如图 1 所示。

图 1：指数转换函数的 ListFold 的概率解释



资料来源: Quantitative Finance，国信证券研究所整理

我们之前构建的损失函数不是对 ListMLE 的一个简单概括。实际上，一个简单的概括可能是:

$$\begin{aligned}\mathbb{P}_t(\pi|X,f) &= \prod_{i=1}^{n}\frac{\psi(f_i)}{\sum_{i\le u\le 2n}\psi(f_u)}\\ &\quad\times\prod_{j=1}^{n}\frac{\psi(-f_{2n+1-j})}{\sum_{j\le u\le 2n}\psi(-f_{2n+1-u})},\end{aligned}\tag{7}$$

这更像是通过反向标记来组合两个 ListMLE。这种代理损失与排列级别的 0-1 损失天然保持一致性。但是P_t并没有定义排列空间上的概率，而我们的推广P_c暗示了一个ψ是指数的概率模型。

3.2 理论分析

在这一部分中，我们讨论了我们的损失函数$\mathcal{L}_c$和 0-1 损失的一致性。当ψ是 s 型或指数型时，我们将$\mathcal{L}_c$分别表示为$\mathcal{L}_c^s$ 和$\mathcal{L}_c^e$。在开始之前，我们首先花一些精力来理解为什么这是一个新的具有挑战性的问题。在 ListMLE 之前的工作中，代理损失函数都是顺序敏感的，基本上是说如果我们将任意两个文档的位置向 ground truth 交换，损失就会减少。order sensitive 的正式定义可以在 Xia et al.(2008)中找到。顺序敏感性是指只要神经网络在任意两点上学习得更好，损失就会减少。但实际上当f有一个更全面的预测会更加令人满意。reduces-if 损失承认在股票市场上预测的排名不会完全正确,因此应该愿意允许代理损失，并探索更复杂的模型。从这个角度来看，我们的损失函数$\mathcal{L}_c^e$是一种异常，可能引发提出替代损失函数的新思路。考虑对四个数字(5,4,1,0)的排列。假设我们从$\mathcal{L}_c^e$(1,5,4,0) = 4.78 开始，如果我们交换前两个数字，损失实际上会增加： $\mathcal{L}_c^e$(5,1,4,0) = 6.65，而基本期望值$\mathcal{L}_c^e$ (5,4,1,0) = 0.65 仍然是最小的。

接下来用两个定理分别来描述$\mathcal{L}_c^s$和$\mathcal{L}_c^e$的一致性。这两个定理试图回答当最小化等式(4)中定义的 ListFold 时，我们的目标函数是什么样的真实损失。理论分析将帮助我们更好地解释模型输出，并据此构建最优的多空组合。

定理 3.1 如果变换函数ψ为 sigmoid，则设$a_1 \geq a_2 \geq \cdots \geq a_n \geq b_n \geq b_{n-1} \geq \cdots \geq b_1$，且$f \coloneqq (f_1, \ldots, f_{2n})$是所有$a_i$和$b_i$的排列。那么我们的损失$\mathcal{L}_c^s(f, y, X)$与二元分类损失一致。

证明：根据定义$sigmoid(x) = 1/(1 + e^{-x})$和$sigmoid(x) + sigmoid(-x) = 1$的性质，

$$\begin{aligned}
&\mathcal{L}_c^e(f, y, X) \\
&= -\sum_{i=1}^{n} \left(\log \psi(f_i - f_{2n+1-i}) - \log \sum_{i \leq u \neq v \leq 2n+1-i} \psi(f_u - f_v) \right) \\
&= -\sum_{i=1}^{n} (\log \text{sigmoid}(f_i - f_{2n+1-i}) + (2n - 2i + 2)) \\
&= \sum_{i=1}^{n} \log(1 + e^{-f_i + f_{2n+1-i}}) + C_n, \qquad (8)
\end{aligned}$$

其中$C_n = n(n+1)$是一个固定的常数 n。由于$log(1 + e^{-x})$是凸函数并且单调递增，对于任何四个分数$f_i \leq f_j \leq f_k \leq f_l$，考虑他们所有的排列$(a, b, c, d) \epsilon Perm(f_i, f_j, f_k, f_l)$，我们取其中的两对，考虑损失函数：

$$\mathcal{L}_c^s(a, b, c, d) = \log(1 + e^{-(a-b)}) + \log(1 + e^{-(c-d)}).$$

为了使损失最小化，需满足$a > b$, $c > d$，否则可以互换头寸来减少损失。因此只需要比较三种情况$f_l - f_i, f_l - f_j, f_l - f_k$:

$$\begin{aligned}
&\log(1 + e^{-(f_l - f_i)}) + \log(1 + e^{-(f_k - f_j)}) \\
&\leq \log(1 + e^{-(f_l - f_j)}) + \log(1 + e^{-(f_k - f_i)}), \\
&\log(1 + e^{-(f_l - f_i)}) + \log(1 + e^{-(f_k - f_j)}) \\
&\leq \log(1 + e^{-(f_l - f_k)}) + \log(1 + e^{-(f_j - f_i)}),
\end{aligned}$$

第一个不等式成立是因为它的凸性，第二个不等式成立是因为它的单调性。对每两对保持使用这一规则，只要排列对在一起$(a_1, b_n), (a_2, b_{n-1}), \ldots, (a_n, b_1)$，损

失是最小的。这 n 对的排列实际上没有区别。因此，$\mathcal{L}_c^s$与二分类损失函数一致，但与 0-1 排列损失函数不一致。

定理 3.2 如果变换函数ψ是指数函数，假设$a_1 \geq a_2 \geq \cdots \geq a_n \geq b_n \geq b_{n-1} \geq \cdots \geq b_1$，且$(f_1, \ldots, f_{2n})$是所有$a_i$的排列，$(f_{-1}, \ldots, f_{-n})$是所有$b_i$的排列，则表示$f := (f_1, \ldots, f_n, f_{-n}, \ldots, f_{-1})$，损失函数为$\mathcal{L}_c^e(f, y, X)$

$$\mathcal{L}_c^e(\mathbf{f}, y, X) = \sum_{i=1}^{n} \left(-(f_i - f_{-i}) + \log \sum_{-i \leq s \neq t \leq i} e^{f_t - f_s} \right)$$

在倒序的排列中可得到最小值$f^* := (a_1, \ldots, a_n, b_n, \ldots, b_1)$。

证明略

从证明中，可以看到有两种方法可以让$\mathcal{L}_c^e$在排列空间中减少：

- 将较高的分数放在头部，将较低的分数放在尾部。
- 将小分数差异（较难区分的）成对放在中间。

第二种方法，一方面给证明排列水平 0-1 一致性带来困难；另一方面，它实际上与股票市场是一致的：股票收益的横断面分布近似正态分布，排在中间的几乎没有差别。从多空的角度来看，对于那些不确定其中一个是否会在回报方面严重影响另一个的股票组合的，将其放在一起。

4. 实证研究

在这一节中，我们探讨了中国 A 股市场的实证表现。将构建的模型与 MLP、ListMLE 和 Song 拟合的两个 ListMLE（记为 List2MLE）进行比较。由于我们重点关注损失函数，所以接下来将采用相同的神经网络结构和训练方法。MLP 代表价值预测，其余的则代表排序预测。MLP 的损失函数为：

$$\mathcal{L}_{MLP}(f, r) = \frac{1}{n} \sum_{i=1}^{n} (r_i - f_i)^2.$$

针对不同的算法，我们构造了相应的多空组合。然后将评价指标作为投资组合绩效的度量。在排序预测方面，还提出了一种广义 NDCG 评价方法。这块分为三个部分：数据和训练过程的简要总结，网络结构的打分函数以及表现的评价。

4.1 数据和训练

我们的数据集包含了中国 A 股市场 3712 只股票的 631 周的数据，其中包含 68 个因子，样本日期为从 2006-12-29 到 2019-04-19。通过筛选缺失值百分比小于 0.1%的样本，我们过滤掉了其中的 80 只股票。这 80 只股票大多是沪深 300 指数成份股，流动性较高。这 68 个因子如下图所示：

图 2：使用到的因子名称

alpha_100w	amount_21	amount_5	amount_63	amount_div	avg_volume_21	avg_volume_5
avg_volume_63	beta_100w	close_low_high	close_s_vwap5	close_vwap5	c_l2_ibm	dlt_mclo
highlow_1	highlow_12	highlow_3	highlow_6	ibm_close	ibm_svlo	IR_netasset_252
IR_roe_252	l2_ibm_ewma	l2_lbm_ewma	magm_yop	ma_crossover_15_36	net_assets	n_buy_value_small_order
pb	pcf_gm	r_sde_pe	q_s_fa_yoyocf	rank_amount_div	rank_close_low_high	rt_10
rt_126	rt_12_1	rt_15	rt_21	rt_252	rt_5	rt_5_Skewness_10
rt_5_Skewness_15	rt_5_Skewness_20	rt_5_Skewness_5	rt_63	std_deviation_100w	yop_pe	s_dq_mv
yop_pcf	s_val_mv	z_rank_pe	trk_rk_pe_re	ttm_pcf	ttm_pe	ttm_ps
ttm_roa	ttm_roe	turnover_21	turnover_5	turnover_63	vol_1	vol_12
vol_3	vol_6	yieldvol_1m	yieldvol_3m	yieldvol_6m		

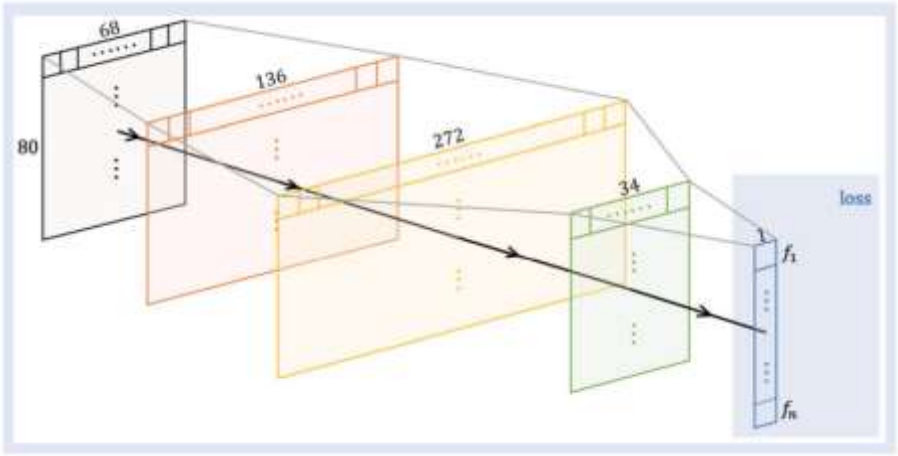
资料来源：Quantitative Finance，国信证券研究所整理

由于数据集的跨度超过 13 年，接下来以滚动的方式训练模型。我们将每 300 周作为训练集，接下来的 16 周作为测试集。我们将数据分成 32 个小批量数据，每 300 周执行一次最小最大归一化。最终得到 320 个周度数据作为从 2012-11-30 到 2019-02-01 的测试集。在每次训练过程中，设定每个算法查看 1000 个小数据集。

4.2 打分函数

我们使用 4 层全连通神经网络学习评分函数f，即网络结构的形状为[68 × 136 × 272 × 34 × 1]。在每层中都嵌入一个 ReLU 层作为激活函数，同时为了这些因子相互作用，以生成更多的特征，我们扩展了隐藏层中的特征的维度。对于所有的样本，打分函数都是相同的，比如$f_1, \ldots, f_n$的参数都相同。如图 3 所示：

图 3：学习打分函数的神经网络



资料来源：Quantitative Finance，国信证券研究所整理

在确保模型不会使用未来信息的基础上，将每周作为独立样本，在时间轴上，神经网络不会生成新的特征，这使我们能够在训练过程中无缝地进行小批量数据的交叉训练。在实践中，只需将这些信息作为时间序列因子输入，就可以将模型推广到沿时间轴的维度。

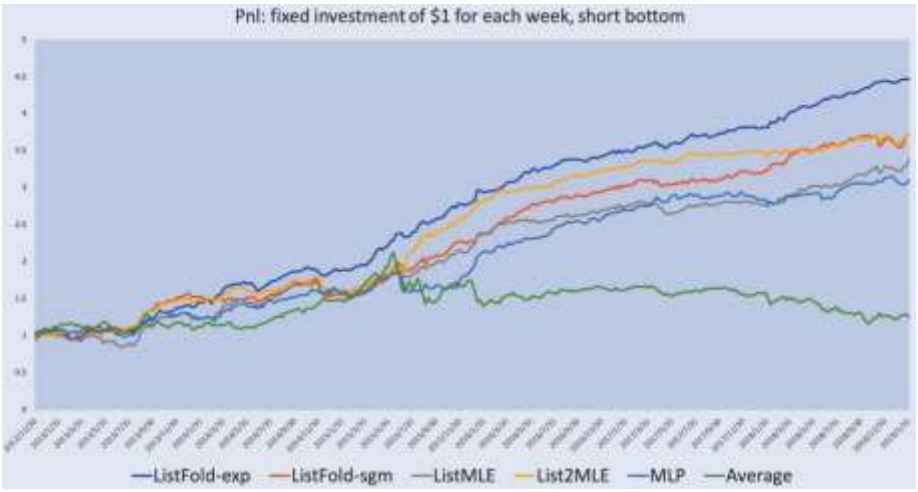
4.3 组合表现

为了评估模型，首先查看投资组合样本外的表现。然后我们从排序的角度进行评价，希望它能帮助我们更好地分解利润，通过 IR 指标来衡量模型的效果。

4.3.1 度量投资组合

在得到排序预测的基础上，我们构建了两种策略。一种是做多头部 10%的股票，同时做空尾部 10%的股票。另一种是做多头部 10%的股票，做空所有股票的平均值。对相同投资方向的股票分配相同的权重。每周投入 1 美元，没有交易费用。样本外净值曲线在图 4 中绘制，以平均值作为基准。

图 4：不同损失函数的多空投资组合净值曲线-做空尾部 10%的股票



资料来源：Quantitative Finance，国信证券研究所整理

图 4 展示了做空尾部 10%的股票模型中排序对投资组合的贡献。在相同的神经网络、相同的数据和相同的训练程序下，回测区间内 ListFold-exp 的收益最高，MLP 的净值仅为 ListFold-exp 净值的 2/3，ListFold-exp 优于 List2MLE。

图 5：不同损失函数的多空投资组合净值曲线-做空所有股票的平均值

资料来源：Quantitative Finance，国信证券研究所整理

图 5 展示了做空所有股票的平均值模型中排序对投资组合的贡献。对比图 4 和图 5，我们发现 ListFoldexp 和 ListFold-sgm 在 long leg 上表现更好，几乎所有 listFold-exp 和 MLP 之间的超额收益都来自 long leg。自 2016 年以来，如果 ListMLE 和 MLP 做空平均水平则收益甚微，而 ListFold 的两种策略仍然能够获得不错的收益。一种可能的解释是，机器学习算法在此之后涌入了中国 A 股市场。

设定年化无风险率r_f为 3%，每笔交易的总交易成本为 30 个基点（如税费、价差交叉和卖空成本），最终得到了平均值、标准差、夏普比率和最大回撤指标。此外还计算了平均交易成交量（简称 TRV），即连续两周持仓的非重叠股票比率。如下图所示，第一块数据显示做空尾部股票的策略，第二块数据对应做空平均线的策略。

图 6：两种策略结果统计

	ListFold-exp	ListFold-sgm	ListMLE	List2MLE	MLP
$\mu - r_f$	0.38	0.26	0.20	0.26	0.16
σ	0.19	0.20	0.22	0.20	0.22
SR	2.01	1.27	0.91	1.29	0.72
MDD	0.14	0.25	0.23	0.21	0.28
TRV	0.48	0.45	0.45	0.46	0.39
	ListFold-exp-sa	ListFold-sgm-sa	ListMLE-sa	List2MLE-sa	MLP-sa
$\mu - r_f$	0.08	0.06	−0.06	×	−0.19
σ	0.11	0.11	0.10	×	0.11
SR	0.71	0.50	−0.53	×	−1.79
MDD	0.09	0.10	0.12	×	0.27

资料来源：Quantitative Finance，国信证券研究所整理

所有这些策略由于其多空性质而具有较低的波动性，其中 ListFold-exp 优于 List2MLE，但其换手率略高。

4.3.2 排序指标

对于排序指标，我们使用 Spearman 的ρ，$NDCG$。我们还提出$NDCG@\pm k$作为$NDCG$的一种概括，它同时强调头部和尾部:

定义 3 $NDCG@\pm k$被定义为$NDCG@k$和反向标记$NDCG@-k$的平均值:

$$\begin{aligned}&\mathrm{NDCG@}\pm k(\pi, l)\\&= (\mathrm{NDCG@k}(\pi, l)\\&\quad + \mathrm{NDCG@}-\mathrm{k}(\pi, \tilde{l}))/2\\&= \frac{1}{2Z_k}\left(\sum_{j=1}^{k} G(l_{\pi^{-1}(j)})\eta(j) + \sum_{j=1}^{k} G(\tilde{l}_{\pi^{-1}(j)})\eta(j)\right),\end{aligned}$$

尽管 LisMLE 具有很高的 NDCG，但它的 NDCG@8 在我们的数据集中并不那么好。List2MLE 则选取了 ListMLE 的 long leg 和 ListMLE-rvs 的 short leg，因此其 NDCG ± 8 实际上是最低的。此外，IC 指标与 pnl 的表现比较接近，也反映了 IC 指标在业界的受欢迎程度。

4.4 鲁棒性

在这一部分中，我们首先检验了仓位截止参数 k 的鲁棒性，即我们决定在投资组合中做多和做空的头部部分 k 和尾部部分 k。对于不同的模型，我们绘制了不同数量股票的多空组合的热力图。列代表不同模型，行代表仓位截止参数 k。

图 7：多空截至参数 k 组合对热力图

	ListFold-exp	ListFold-sgm	ListMLE	List2MLE	MSE
ls-1	105	96	-15	70	-29
ls-2	80	67	44	121	-12
ls-3	96	75	67	115	21
ls-4	101	81	71	105	42
ls-5	112	87	69	92	42
ls-6	110	90	69	93	48
ls-7	106	90	65	84	58
ls-8	108	89	75	86	66
ls-9	105	87	76	80	64
ls-10	101	86	78	82	68
ls-11	98	86	76	76	65
ls-12	89	84	72	72	62
ls-13	86	83	72	73	59
ls-14	81	82	69	70	52
ls-15	79	81	69	71	51
ls-16	78	81	69	68	50
ls-17	78	80	69	66	53
ls-18	76	79	68	66	54
ls-19	72	79	65	65	52
ls-20	69	78	63	65	51
ls-21	67	77	61	63	52
ls-22	68	76	60	61	51
ls-23	66	75	58	58	51
ls-24	64	74	58	57	49
ls-25	62	73	56	55	48
ls-26	62	72	53	54	48
ls-27	61	71	52	53	48
ls-28	62	70	52	51	47
ls-29	61	69	51	51	43
ls-30	60	68	49	50	41
ls-31	58	67	48	47	41
ls-32	58	66	47	47	38
ls-33	57	66	46	45	37
ls-34	54	65	46	45	35
ls-35	53	64	45	43	36
ls-36	52	63	44	42	35
ls-37	51	63	43	41	34
ls-38	50	62	42	39	35
ls-39	49	61	41	38	35
ls-40	48	60	39	37	35

资料来源：Quantitative Finance，国信研究所整理

例如，在第 8 行中，我们做多头部的 8 只股票，做空尾部的 8 只股票。每个单元格中的数字是在没有交易费用的情况下从样本中获得的平均周回报（bps）。这个值越大，颜色就越靠近暖色调。

如果我们的预测完全符合 ground truth，可以预期，当 k 从 1 到 80 时，投资组合的 PNL 逐渐向 0 下降。图 7 生动地展示了 ListFold-exp 的优势。四种排序方法都对 k 具有鲁棒性。如果 LS-1 到 LS-8 的值近似减小，我们也可以通过给先例对分配更大的权重来提升 pnl。除此之外，我们还注意到 ListFold-sgm 在所有五种车型中长-短-40 性能最好，性能稳定。这与 ListFold-sgm 与二进制分类丢失是一致的事实是一致的。因此，如果一个人的任务是做多 50%，做空 50%，

那么 ListFold-sgm 应该是一个不错的选择。

接下来，我们研究了小批量排序方法的稳健性。我们在不同的小批量尺寸下训练模型，并保持模型始终查看 1000 批。表 4 总结了以 bps 计算的平均周收益。总的来说，小批量的小批量不适合 ListFoldexp 和 ListFold-sgm，因为它不是凸的。虽然一个更复杂的损失函数通常需要更多的数据，但早期停止规则也是必要的。

最后，我们想提到的是，我们的策略不限于每周的方式或 80 支股票。它可以无缝地传输到从业者自己的因素数据集。我们模型的高夏普比率也鼓励在实际生产中使用杠杆。由于中国 a 股市场在我们的测试期经历了一个非同寻常的牛市-熊市循环，我们的策略已经证明了其对市场动荡的稳定期。

5. 结论

本文提出了组合因子构建多空组合的新思路。在学习排序方法的基础上，我们提出了一种新的损失函数，目的是通过 listwise 方式选择多空对。它具有平移不变和概率可解释的特点。对于不同的变换函数，它和二分类损失函数或排列水平的 0-1 损失函数相一致。我们的模型可以被看作是研究非顺序敏感损失函数的一个补充工具，它也可以启发统一 pairwise 和 listwise 代理损失函数的框架。

我们对中国 A 股市场的实证研究结果表明，6 年来年化收益率达到 38%，夏普比率为 2。它不仅展示了排序预测相对于数值预测的优势，而且还展示了我们的损失函数，特别是 ListFold-exp 的强大作用。我们从财务和排名的角度对不同模型进行了全面的评估。结果表明，我们提出的损失函数比其他函数有明显的优势。本文研究的一个副产品是，我们通过经验验证了 IC 比 NDCG 类型的排名指标更适合评估 alpha 策略。

在未来的研究中，特别是在理论分析方面，值得进一步研究 ListFold-exp 的一致性，更一般地，对导致不同实际损失函数的不同变换函数进行刻画。以从业者的角度来看，在他们自己的因子库或神经网络模型上尝试 ListFold 也是值得的。事实上，考虑到做空通常更昂贵的事实，一个人可能希望尾部排序的正确率相对于平均更高。因此，将 ListFold 与其他损失函数结合起来，或者在当前架构中添加神经网络模块（如辅助任务），可能也是很有意思的领域。

分析师承诺

作者保证报告所采用的数据均来自合规渠道，分析逻辑基于本人的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

风险提示

个股投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。

证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信个股经济研究所

深圳
深圳市罗湖区红岭中路 1012 号国信个股大厦 18 层
邮编：518001　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 楼
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信个股 9 层
邮编：100032