如何使用强化学习优化动态资产配置？
——“学海拾珠”系列之一百七十九

报告日期：2024-2-21

分析师：严佳炜
执业证书号：S0010520070001
邮箱：yanjw@hazq.com

分析师：吴正宇
执业证书号：S0010522090001
邮箱：wuzy@hazq.com

相关报告
1. 《如何改进短期反转策略？——“学海拾珠”系列之一百七十》
2. 《如何衡量基金产品创新与差异化——“学海拾珠”系列之一百七十一》
3. 《低风险组合构建：基于下行风险的缩放策略——“学海拾珠”系列之一百七十二》
4. 《基于端到端神经网络的风险预算与组合优化——“学海拾珠”系列之一百七十三》
5. 《历史持仓回报会影响基金经理后续选股吗？——“学海拾珠”系列之一百七十四》
6. 《基于残差因子分布预测的投资组合优化——“学海拾珠”系列之一百七十五》
7. 《美元 beta 与股票回报——“学海拾珠”系列之一百七十六》
8. 《基金经理技能之卖出能力的重要性——“学海拾珠”系列之一百七十七》
9. 《高成交量回报溢价与经济基本面——“学海拾珠”系列之一百七十八》

主要观点：

本篇是“学海拾珠”系列第一百七十九篇，本文研究如何使用强化学习进行动态资产配置，特别关注**投资约束和非平稳性**方面的影响。作者首先探讨了**将金融时间序列数据的非平稳性纳入强化学习算法**的重要性，研究结果强调了在环境设置中引入如状态(regime)变化等特定变量以提高预测准确性的重要性。此外，强化学习在配置策略中的优化条件设置上具有显著优势，使得可以将投资者面临的实际约束集成到算法中，从而实现有效的优化。

回到国内市场，机器学习和量化投资的结合当前大多应用于选股，资产配置领域的研究相对较少，本文对强化学习和动态资产配置的研究值得学习。

- **非平稳性与强化学习**

本研究强调了在强化学习模型中纳入金融时间序列数据的非平稳性的重要性，通过考虑市场状态变化等变量，显著提升了模型的预测性能。这表明，通过对市场动态和结构性变化的深入理解，强化学习模型能够更准确地捕捉到市场机会。

- **投资约束的集成**

作者成功地将投资过程中遇到的各种实际约束集成到强化学习框架中，包括风险管理、资金要求和交易成本等，展示了在满足这些约束条件下如何进行有效的资产配置。这一发现突出了强化学习在处理复杂投资决策问题时的灵活性和应用潜力。

- **通过强化学习优化动态资产配置**

通过实证分析，本文展示了强化学习在动态调整资产配置以适应市场变化中的有效性。特别是，研究通过考虑市场的非平稳性和实际投资约束，提出了一种能够在多变市场环境下优化投资组合表现的方法。

- **风险提示**

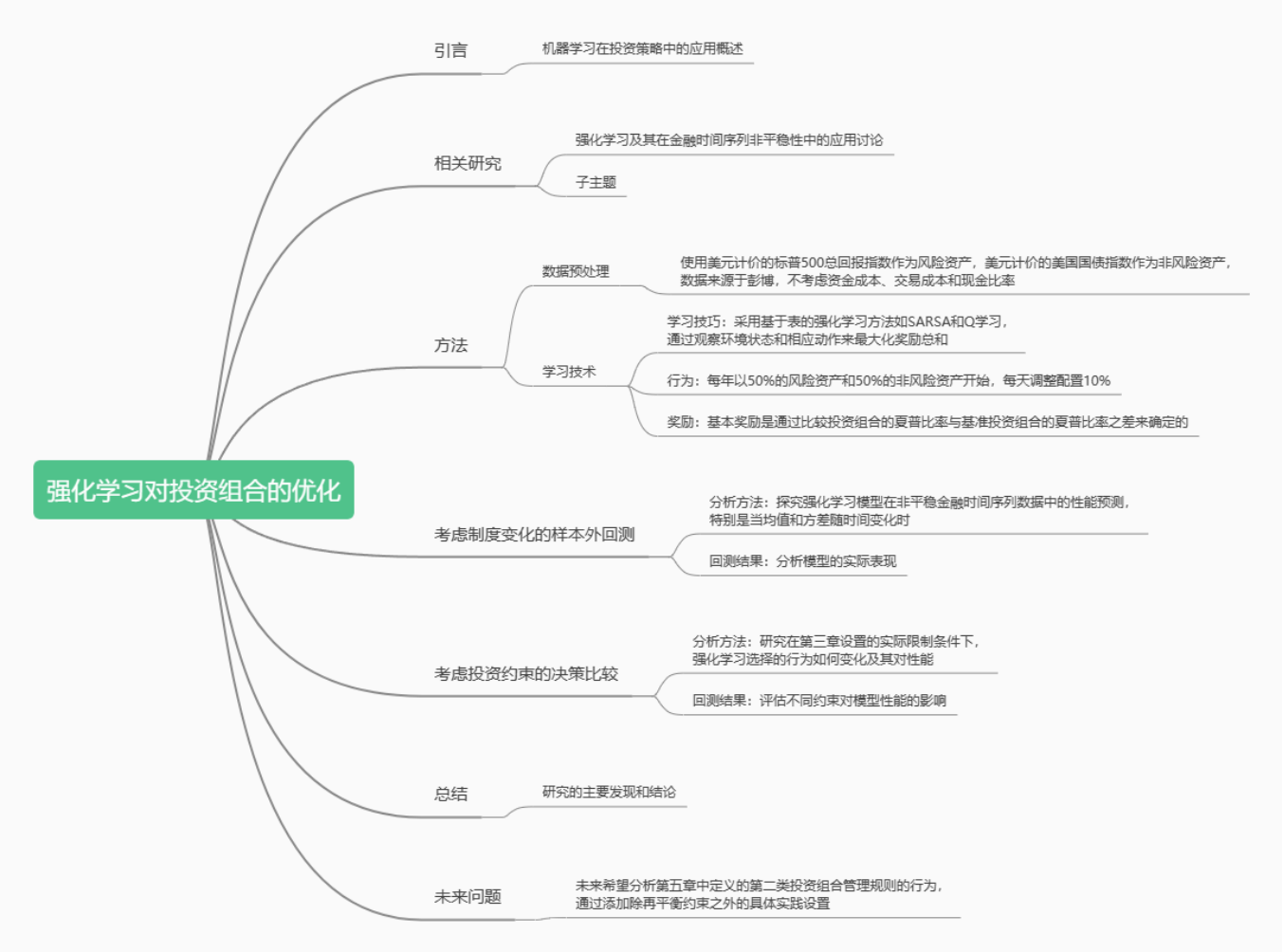
文献结论基于历史数据与海外文献进行总结；不构成任何投资建议。

正文目录

图表目录

1 引言

图表 1 文章框架



资料来源：华安证券研究所整理

近年来，机器学习应用于投资策略的研究和开发取得了显著进展。这些研究方向主要分为两类：**第一类**是使用深度学习构建的模型来阅读或解释非结构化数据，如文本和图像；**第二类**是使用机器学习模型估计先前通过财务工程中的线性或统计模型估计的参数。后者的优势在于能扩展到非线性和更复杂的模型，如深度学习和集成机器学习，从而提供比传统模型更高维度上估计参数的可能性，因此可以构建操作性能更高的模型。投资组合经理使用典型量化方法的投资决策过程可以分为三个步骤：数据收集与处理、分析与信号生成以及**投资组合优化**。本研究聚焦于在投资决策制定阶段中如何应用机器学习和人工智能技术，即第三步。在这种情况下，基于对财务会计、法律和监管事务以及税务的多方面理解，寻找公司的最佳投资组合变得必要，而不仅是追求理论上的高效投资组合。因此，本研究考察了强化学习方法在制定包含此类实际约束的最优投资组合中的应用价值。

2 相关研究

强化学习是机器学习的一部分，设计用于学习顺序决策规则。强化学习的独特之处在于，它**通过奖励来达成目标，并且在不完全了解其应用的系统或环境的情况**

下，从数据中学习如何实现这些目标。在投资决策中，使用强化学习主要考虑的是如何建模金融时间序列数据的非平稳性。这意味着均值、方差和协方差随时间不是常数，历史上曾观察到诸如阶段突变或波动率激增等多个事件。在强化学习中应对非平稳性的特定方法已在以往的文献中描述。

3 研究方法

3.1 数据预处理

本研究检验了通过决策重新平衡两种资产，即风险资产和无风险资产，来实现动态最优配置。使用的数据是从 2000 年 4 月到 2023 年 3 月的每日数据。作者选择以美元计价的标普 500 总回报指数作为风险资产，以美元计价的美国国债指数总回报指数作为无风险资产，不考虑资金成本、交易成本和现金比率。

3.2 学习技术

（1）学习技术

与第 4 章和第 5 章一样，本研究中强化学习模型学习的过程如下。强化学习涉及通过观察环境中的状态及其相应的行动来最大化奖励总和。为了分析每个状态的行为，使用基于表的强化学习方法，如 SARSA 和 Q 学习。

<SARSA 更新表达式>

$$Q'\pi\,(St\,,At\,)=Q\pi\,(St\,,At\,)+\alpha\{Rt+\gamma Q\pi\,(St{+}1,At{+}1\,)-Q\pi(St\,,At)\}$$

<Q Learning 更新表达式>

$$\mathrm{Q}'\,(St\,,At\,)=\mathrm{Q}(St\,,At\,)+\alpha\,\{Rt+\gamma\max a\,Q(St{+}1,a)-Q(St\,,At)\}$$

这里的 Q 代表行动价值，R 是即时奖励，S 是状态，A 是行动，α 是学习率，γ 是折现率。ε-贪婪方法被用来在每个研究周期中研究 1000 个

(2) 行动

每年年初，风险资产和无风险资产各占 50%，配置会以 10%的比例变动。有三种可能的行动：[风险资产：+10%，无风险资产：-10%]、[权重不变]、[风险资产：-10%，无风险资产：+10%]。然而，如果权重已达到 100%，即使选择增加权重，权重也不会变动。

(3) 奖励

基本奖励是**通过计算投资组合的夏普比率与基准投资组合的夏普比率之间的差值**（包括年初至今和过去 10 天），然后将这些差值相加来确定的。对于基准投资组合，每年对于风险资产和无风险资产都固定权重的最高配置的夏普比率被确定。这种配置被视为回顾性确定的正确配置。

夏普比率的计算如下：

$$SR=\frac{r}{\sigma}$$

其中r代表投资组合回报率，σ代表投资组合回报的标准差。无风险利率为 0。时间 t 的基本补偿由以下公式给出：

$$R_t^{Base}=(SR_t-SR_t^{BM})+(SR_t^{10}-SSR_t^{BM,10})$$

SR_t^{BM}是基准投资组合的夏普比率，SR_t^{10}是根据过去 10 天的回报计算的夏普比率。在每次分析中，可能会向基本奖励添加额外的奖励，这将在后续详细描述。

(4) 状态

每个分析案例的状态不同，将在后文详细描述。

4 考虑状态变化的样本外回测

4.1 分析方法

本章探讨了强化学习模型用于预测时，金融时间序列数据的非平稳性，特别是当均值和方差随时间变化时，如何影响预测能力。作者考虑了两种模型。第一种模型仅以风险资产和无风险资产的预期回报来定义状态空间。对于预期回报，简单地使用 60 日前的价格差异（动量）。基于资产动量的正负进行**二元分类**，并将状态空间划分为 2x2 的四个状态。第二种模型在状态变量中加入了风险资产和无风险资产之间的相关系数。相关系数是根据过去 60 日的日数据估计的。相关系数状态变量被划分为**正相关/无相关/负相关三个状态**，阈值为±0.2，并与预期回报结合，总共形成 12 个状态。本章使用 SARSA。第一个模型被称为基础模型，第二个模型被称为非平稳模型，并且与使用 Q 表作为随机变量的随机模型的回测评估也进行了比较。随机模型通过 1000 次生成随机 Q 表来衡量表现。

4.2 回测结果

在学习期间，每个财年的 4 月到次年 3 月被用作一组，以估计每组的 Q 表。对于样本外回测验证，使用过去学习期间而非验证期间估计的 Q 表，并且未来期间的 Q 表不被应用。例如，在进行 2018 年 4 月到 2019 年 3 月的样本外回测验证时，计算 2000 财年到 2017 财年研究期间估计的总共 18 个 Q 表的每个元素的平均值，使用等权重。此外，再平衡频率为每日，且不考虑交易成本。表 2 比较了每年的样本外表现与基础模型、非平稳模型和随机模型的中位数。回测表现由夏普比率定义，即年化回报除以年化标准差。从 2001 年到 2022 年，**非平稳模型的平均夏普比率高于基础模型，差异具有统计学意义**。非平稳模型的平均夏普比率也优于随机模型的中位数。比较每年的结果，回测期间的前半部分，即 2000 年代，非平稳模型表现不如基础模型的年份比后半部分，即 2010 年代的更优。这可能是因为学习期从 2000 财年开始，未提供足够的时间来预测未来。

图 3 比较了随机模型的夏普比率的概率密度与随机 Q 表的概率密度。这项分析的目的是假设某些年份中，资产回报的表现显著取决于其短期动量和年内日变化率分布的形状，而某些年份则不然。例如，在发生重大事件（如总统选举或 FOMC 会议期间），随机模型的概率密度被知为双峰，且年度表现根据该时期投资组合是否持有大量风险资产而发生戏剧性变化。

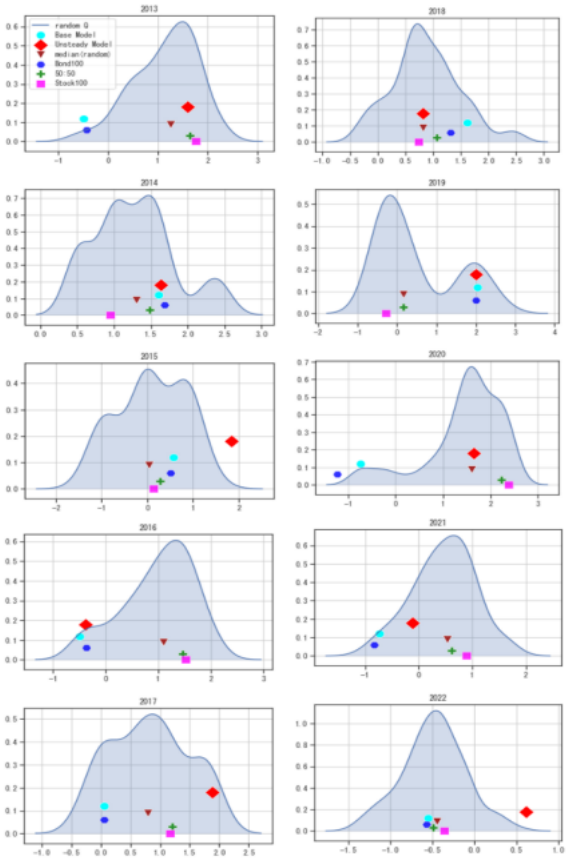
图表 2 回测的历年夏普比率

FY	Base Model	Random Model (Median)	Non-Stationary Model
2001	0.42	0.16	0.59
2002	2.21	-0.69	1.38
2003	0.97	1.75	0.36
2004	0.33	0.48	-0.01
2005	0.62	0.95	1.28
2006	2.05	1.15	2.03
2007	2.34	0.19	2.74
2008	1.08	-0.01	1.08
2009	0.01	1.63	0.01
2010	1.01	0.77	1.16
2011	1.72	0.79	1.72
2012	0.85	1.01	0.85
2013	-0.49	1.25	1.59
2014	1.61	1.30	1.64
2015	0.57	0.03	1.85
2016	-0.49	1.10	-0.37
2017	0.06	0.79	1.88
2018	1.62	0.81	0.81
2019	2.04	0.16	2.00
2020	-0.74	1.60	1.65
2021	-0.73	0.53	-0.12
2022	-0.56	-0.45	0.61
Average(All)	0.75	0.70	1.12
Average(2005～)	0.70	0.76	1.24

资料来源：《 Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning 》，华安证券研究所

在这种情况下，金融时间序列数据的非平稳性得以显现，证实了非平稳模型的有效性。如图 3 所示，观察到非平稳模型的夏普比率在超过一半的年份中超过了随机模型的中位夏普比率。此外，确认在具有双峰形态的年份中，非平稳模型在夏普比率较低峰值范围内的年份较少。这些结果表明，向强化学习模型的状态变量添加阶段有助于显著提高预测准确性。

图表 3 近十年的结果（横坐标：夏普比率）



*Bond100 = 100% non-risky assets, 50:50 = 50% risky assets: 50% non-risky assets, Stock100 = 100% risky assets Buy & hold for 1 year

资料来源：《 Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning 》，华安证券研究所

5 考虑投资约束的决策比较

5.1 分析方法

在本章中，作者探讨在考虑第 3 章设定的实际限制时，强化学习中选择的行为如何转变并影响模型表现，而前一章关注于预测效果，本章比较了样本内学习的结果，以分析不同约束与决策制定之间的关系。下图展示了在每个时间点优化投资组合时需要考虑的三类指标和规则。第一类可测量的绩效指标。例如，除了财年的目标回报和夏普比率外，还使用了如 VaR（风险价值）和回撤等风险指标，以及与投资组合风险回报相关的管理指标，如止损点。第二类包括与投资组合管理相关的规则。例如，与**投资期限和再平衡频率**等周期相关的规则，与结算相关的规定如保证金和清算，与财务指标如杠杆比率、风险加权资产和流动性比率相关的规定，以及与各种金融监管机构相关的规定，如沃尔克规则。第三类包括可以考虑的约束，如预期回报的准确性和交易成本。

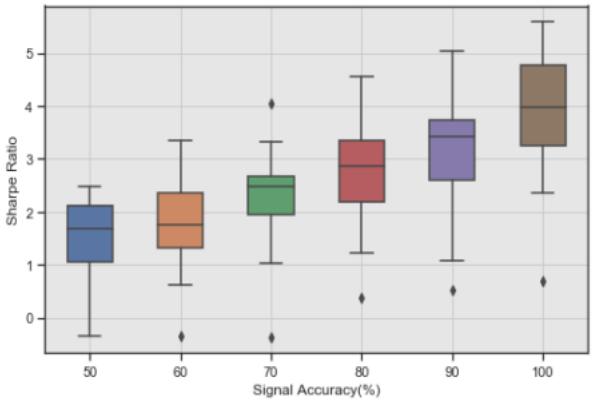
图表 4 投资组合优化中考虑的指标+和规则的例子

Category	Example
① Performance Management Indicators	Target return, target volatility, risk indicators (VaR, drawdown, etc.), loss cut point, etc.
②Portfolio Management Rules	Period, Settlement, Finance, Regulation, etc.
③Other Constraints	Information sources, expected return accuracy, transaction costs, etc.

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

在本章中，作者构建了一个基准投资组合，并将其表现与每个约束下的学习性能进行比较。在进行回测之前，作者评估了由于信号准确性（如图 4 中的第 3 类所示）导致的性能和行为变化。使用未来五个工作日风险资产或无风险资产的夏普比率较高的答案生成二元信号，并据此设置状态。此外，作者还探讨了一定比例的信号被反转（即，用错误的信号替代）的情况。反转率按 10%的步长分类，并检查六种模式，直到真阳性：假=50:50，以确定性能变化。图 5 描述了信号准确性和性能。如前几节所述，分析了 2000 财年到 2022 财年的每一年，并在样本内绘制了每年获得的夏普比率分布。随着水平轴向右移动（信号准确性增加），中位夏普比率单调增加，且随着信号准确性接近 100%，分布宽度扩大。这意味着，**随着信息的优势和信号准确性的增加，预期获得的夏普比率可能会非线性增长**。

图表 5 信号精度和性能



资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

5.2 回测结果

首先，根据图 4 中显示的第一类，设置了七种类型的回测，如下图所示，并比较了行为的变化。在本章中，使用 Q 学习。

图表 6 类别(1)的回测设置绩效管理指标列表

No.	state variables	rewards
#001	signal, position, age	basic reward
#002	Signal, Position, Quarterly, Target Achievement Status (1 step)	basic reward + Target Achievement Reward (1 level)
#003	Signals, Position, Quarterly, Target Achievement Status (2 levels)	basic reward + Target Achievement Reward (2 levels)
#004	Signal, Position, Quarterly, DD Occurrence (1 level)	basic reward +DD penalty (one level)
#005	Signal, Position, Quarterly, DD Occurrence (2 levels)	basic reward +DD penalty (2 levels)
#006	Signal, Position, Quarterly, Target Achievement Status (1 step), DD Occurrence Status (1 step)	basic reward + Target achievement reward (one level), +DD penalty (one level)
#007	Signal, Position, Quarterly, Target Achievement Status (2 Stages), DD Occurrence Status (2 Stages)	basic reward + Target Achievement Reward (2 levels) +DD penalty (2 levels)

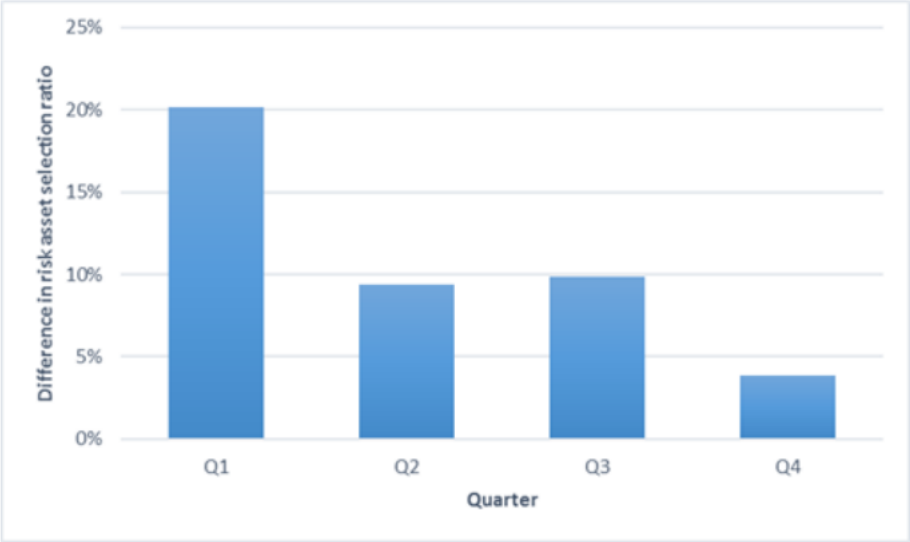
资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

作者使用二进制信号来确定观察到的风险资产或无风险资产的夏普比率哪个表现更好。在信号有效性变化的概率情况下，使用 60%的准确性信号来考察强化学习学到的行为。根据当前持有比率，生成三种配置：高比例的风险资产/等比例的无风险资产/高比例的无风险资产。这是因为“不改变权重”的行为的意义取决于当前权重相对更高的资产。将一年分为四个相等的部分或季度作为经过的时间段，并创建了四个条件。

作者分析了学习结果的选定行动。当风险资产比率较高时，选择[不改变权重]作为偏好风险资产的行为，当无风险资产比率较高时，则作为偏好无风险资产的行为（在等权重状态选择不改变权重时不包括计算）。作者合并了所有 23 年的结果，以确定在每种条件下选择哪些行为及其相应的比率。

基本情况定义为回测编号#001，并主要考察与基本情况的差异。图 7 展示了当信号显示风险资产表现更好时选择的风险资产的百分比，并减去信号表明无风险资产表现更好时选择的风险资产的百分比。

图表 7 信号和行为



资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

在所有季度中，当信号指示风险资产的表现时，强化学习算法更倾向于积累风险资产，表明它根据信号进行进展。此外，随着时间的推移，差异趋于减小。接下来，通过与基础案例的比较，作者检查了#002 结果的比率。在状态变量中，如果周期开始的累积回报超过 5%，则定义为二进制值的目标达成状态，并且如果累积回报从周期开始超过 5%，则增加额外奖励(+1)。分析与基础案例的行动率差异显示，在第三和第四季度，当目标达成且信号不指示风险资产的表现时，风险资产的积累率减少。如果保持这种状态，他们将获得额外奖励。因此，他们可能学会了维持这种状态以避免风险，而不是通过返回到未达到目标的状态来承担额外风险，是更可取的。此外，当目标未达成时，特别是在第四季度，选择风险资产的比例增加。这表明，**为了获得实现目标的额外奖励，他们学会偏好具有高波动性的风险资产，即大价格范围，以在有限的剩余时间内实现目标**。

图表 8 风险和资产偏好的变化 #002

Signal: Risky assets outperform

	otherwise	target(5%) achieved
Q1	3%	6%
Q2	15%	8%
Q3	3%	-3%
Q4	14%	10%

Signal: Non-risky assets outperform

	otherwise	target(5%) achieved
Q1	17%	23%
Q2	8%	9%
Q3	12%	-4%
Q4	21%	-1%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

在#003 中，作者检查了在两个阶段定义目标的情况。对于状态变量和奖励，设定了一个两阶段目标，从周期开始的累积回报为 5%和 10%，分别增加+1 和+2 的额外奖励。分析与基础案例的行为差异显示，即使在达到第一阶段目标后，选择

风险资产的比例也增加了，与#002 相比。这表明，通过在达成目标后提供形式为额外奖励的激励，可以在不降低风险偏好的情况下做出决策。

图表 9 风险和资产偏好的变化 #003

Signal: Risky assets outperform

	otherwise	target(5%) achieved	target(10%) achieved
Q1	5%	-1%	-5%
Q2	8%	6%	5%
Q3	15%	0%	0%
Q4	13%	19%	14%

Signal: Non-risky assets outperform

	otherwise	target(5%) achieved	target(10%) achieved
Q1	19%	22%	20%
Q2	11%	11%	31%
Q3	9%	4%	-2%
Q4	22%	25%	9%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

#004 中，作者讨论了回撤情况。类似于#002，测算自期初起的累积回报，并且添加了两个值到状态变量，指示回撤是否超过 5%。每当超过时，受到惩罚(-1)。从与基础案例的行动率比较来看，当回撤不发生时，即使它指示风险资产的表现，选择增加风险资产的比例下降。这种现象在回测期间的早期到中期部分尤为明显。这种变化可能是由于**向避险决策转变以避免回撤惩罚**。

图表 10 风险和资产偏好的变化 #004

Signal: Risky assets outperform

	otherwise	drawdown exceeded 5%
Q1	-15%	-1%
Q2	-3%	15%
Q3	-18%	-18%
Q4	6%	5%

Signal: Non-risky assets outperform

	otherwise	drawdown exceeded 5%
Q1	16%	28%
Q2	16%	4%
Q3	-4%	-4%
Q4	7%	2%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

在#005 中，作者分析了设定两个级别的回撤情况。与#002 一样，添加了三个值到状态变量，指示回撤是否超过 5%和 10%。此外，奖励的惩罚也设定为两个级别。额外的惩罚为 5%的-1 和 10%的-2。与基础案例相比，选择风险资产的比例在第二次回撤之后增加，不管信号如何。这表明，当组合表现显著恶化时，决策倾向于偏好具有高波动性的风险资产以避免惩罚。

图表 11 风险和资产偏好的变化 #005

Signal: Risky assets outperform

	otherwise	drawdown exceeded 5%	drawdown exceeded 10%
Q1	-7%	-1%	2%
Q2	-5%	-8%	9%
Q3	-15%	-11%	-6%
Q4	3%	6%	11%

Signal: Non-risky assets outperform

	otherwise	drawdown exceeded 5%	drawdown exceeded 10%
Q1	7%	19%	16%
Q2	0%	10%	16%
Q3	-12%	-9%	0%
Q4	8%	2%	16%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

#006 中，考虑了目标达成状态和回撤发生状态。分析信号指示风险资产的表现，目标未达成，且没有发生回撤的条件显示，在回测期间的前半部分，选择风险资产的比例低于基础模型，而在第二半部分增加。这表明在某一时期内，避免惩罚是优先的，导致从期初开始就偏好避险决策。然而，在周期的后半部，优先考虑实现目标，导致转向风险偏好决策。

图表 12 风险和资产偏好比率的变化 #006

Signal: Risky assets outperform

	target(5%) achieved	otherwise	drawdown exceeded 5%
Q1	-8%	-3%	2%
Q2	-2%	-6%	9%
Q3	1%	7%	6%
Q4	10%	30%	21%

Signal: Non-risky assets outperform

	target(5%) achieved	otherwise	drawdown exceeded 5%
Q1	14%	14%	19%
Q2	20%	4%	10%
Q3	-7%	16%	9%
Q4	-5%	39%	20%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

最后，在#007 中，作者为目标达成状态和回撤发生状态设定了两个级别。

图表 13 风险和资产偏好比率的变化 #007

Signal: Risky assets outperform

	target(10%) achieved	target(5%) achieved	otherwise	drawdown exceeded 5%	drawdown exceeded 10%
Q1	4%	-8%	-2%	2%	6%
Q2	10%	-4%	4%	8%	9%
Q3	-11%	-14%	7%	6%	-15%
Q4	3%	15%	14%	15%	8%

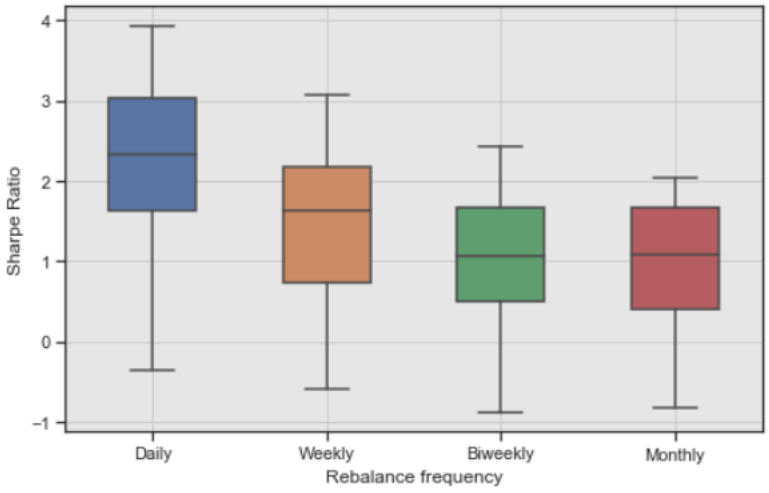
Signal: Non-risky assets outperform

	target(10%) achieved	target(5%) achieved	otherwise	drawdown exceeded 5%	drawdown exceeded 10%
Q1	19%	19%	10%	15%	23%
Q2	9%	7%	13%	15%	11%
Q3	3%	-6%	15%	0%	-13%
Q4	-4%	2%	25%	11%	13%

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

接下来，作者比较了图 4 中展示的第二类，重点关注作为最简单约束的再平衡频率。比较了四种回测场景：每日、每周、每两周和每月再平衡，结果显示，无约束条件展示了最高的平均夏普比率，且**随着间隔的扩大，表现逐渐下降**。

图表 14 再平衡约束



资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

随后，作者分析了信号准确性对于第三类决策制定的影响。类似于#002，作者改变了信号准确性并比较了选择的行为。结果表明，随着信号准确性的提高，选择无风险资产的比率增加，而对于风险资产，尽管趋势比无风险资产弱，但可以按照准确性的顺序辨识出来。

图表 15 信号精度和资产选择比率的变化

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

最后，作者考察了信号准确性在不同阶段变化的情况。数据被分为两个阶段：一个阶段一致使用 60%的准确性信号，而另一个阶段使用从 100%到 50%变化的准确性，这被添加到状态变量中。结果表明，即使没有明确给出信号准确性的差异，个体也能够学会如何有信心地采取适当的行动。

图表 16 信号精度差和资产选择比

资料来源：《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》，华安证券研究所

6 结论

本研究探讨了**如何使用强化学习优化动态资产配置**。从这项研究中得出以下结论：首先，将与阶段变化相关的变量纳入强化学习的状态空间可以增强样本外表现。其次，通过将实际约束纳入强化学习的条件和奖励中，可以改变决策制定。更多的细节如下：

1、在周期内达成目标时，期末更倾向于采取避险行为。

2、相反，如果目标未完成，期末更倾向于采取积极的风险承担。
3、1 和 2 可以在未明确给出每个资产的波动率下达成。
4、即使在两阶段目标的第一阶段目标达成的情况下，仍保留为获取额外回报而倾向于做出承担风险的决策。
5、强制执行回撤阈值并施加惩罚会导致为避免超过阈值而倾向于避险的行为。
6、在两阶段回撤阈值的情景下，如果在第二阶段超过了阈值，将承担更多风险以修复业绩。
7、定义了目标和回撤的情况下，倾向于避免在周期的前半部超过回撤阈值，而采取避险决策，在后半部为实现目标而变得积极承担风险。
8、再平衡频率之间的间隔时间越长，表现越差。
9、信号准确性越低，决策制定变越保守。如果信号准确性在不同阶段变化，将学习哪个阶段的准确性最高，也会增加在准确性高的阶段采取更积极的配置决策。

文献来源：

核心内容摘选自 Yasuhiro Nakayama , Tomochika Sawaki 在 eprint arXiv 上的文章《Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning》

风险提示：

文献结论基于历史数据与海外文献进行总结；不构成任何投资建议。

重要声明

分析师声明

本报告署名分析师具有 PRC 证券业协会授予的证券投资咨询执业资格，以勤勉的执业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，本报告所采用的数据和信息均来自市场公开信息，本人对这些信息的准确性或完整性不做任何保证，也不保证所包含的信息和建议不会发生任何变更。报告中的信息和意见仅供参考。本人过去不曾与、现在不与、未来也将不会因本报告中的具体推荐意见或观点而直接或间接收任何形式的补偿，分析结论不受任何第三方的授意或影响，特此声明。

免责声明

华安证券股份有限公司经 PRC 证券监督管理委员会批准，已具备证券投资咨询业务资格。本报告中的信息均来源于合规途径，华安证券研究所力求准确、可靠，但对这些信息的准确性及完整性均不做任何保证。在任何情况下，本报告中的信息或表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资回报，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。华安证券及其所属关联机构可能会持股报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

投资评级说明

以本报告发布之日起 6 个月内，证券（或行业指数）相对于同期沪深 300 指数的涨回撤为标准，定义如下：

行业评级体系

增持—未来 6 个月的投资回报领先沪深 300 指数 5%以上；

中性—未来 6 个月的投资回报与沪深 300 指数的变动幅度相差-5%至 5%；

减持—未来 6 个月的投资回报落后沪深 300 指数 5%以上；

公司评级体系

买入—未来 6-12 个月的投资回报领先市场基准指数 15%以上；

增持—未来 6-12 个月的投资回报领先市场基准指数 5%至 15%；

中性—未来 6-12 个月的投资回报与市场基准指数的变动幅度相差-5%至 5%；

减持—未来 6-12 个月的投资回报落后市场基准指数 5%至；

卖出—未来 6-12 个月的投资回报落后市场基准指数 15%以上；

无评级—因无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使无法给出明确的投资评级。市场基准指数为沪深 300 指数。