分类号 _____ 密级 _____
U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于深度学习的股票套利策略研究

研 究 生 姓 名：胡娜

指导教师姓名、职称：韩海波、副教授

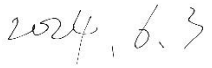
学 科、专 业 名 称：应用经济学、数量经济学

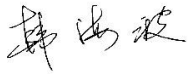
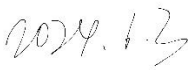
研 究 方 向：金融计量与量化交易

提 交 日 期：2024 年 6 月 5 日

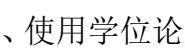
独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：  签字日期： 

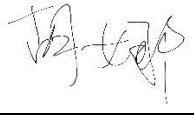
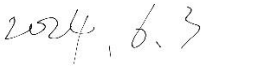
导师签名：  签字日期： 

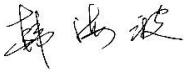
关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，  （选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：  签字日期： 

导师签名：  签字日期： 

Research on stock arbitrage strategy based on deep learning

Candidate : Na Hu

Supervisor: Haibo Han

摘 要

近年来，我国股票证券市场开始蓬勃发展，套利策略逐渐在量化交易的发展中实现从理论研究到制度设计再到实践探索的基本过程，是我国证券市场发展的重要里程碑。随着金融市场交易数据的指数爆炸式增长，影响股票价格的因素也越来越多，因此，对分析金融市场技术工具的要求也越来越高。计算机技术的发展使得机器学习方法进入大众视野，深度学习技术是机器学习中的重要组成部分，在处理非线性关系和高维数据上都有着巨大的优势。

长短期记忆网络模型（LSTM）因其出色的递推性质，适合处理时间序列相关的问题。但 LSTM 模型具有其独特的结构属性，这并不能从根本上解决长期依赖性问题。Transformer 模型基于自注意力机制，并加入了并行注意力机制，这使得它能够捕获远距离的时序特性，避免了长期依赖。因此，本文结合 LSTM 和 Transformer 的模型特点，构建 trans_LSTM 融合模型，并设计了一个基于深度学习的股票套利策略，进一步与 LSTM 模型和统计方法协整模型进行了详细的比较分析。本文第一部分概述了研究的背景意义，概述了国内外的研究进展，并对本文的研究内容和创新点做了简洁的描述。第二部分对套利的基本知识进行了简单介绍。第三部分对基于 trans_LSTM 的套利模型设计进行模型和策略的介绍。第四部分实证分析，通过遍历股票组，基于 trans_LSTM 模型、LSTM 以及协整理论构建预测模型，对套利组合的价差进行预测。

通过构建完整的 trans_LSTM 融合模型和 LSTM 交易策略的套利模型，就科创板数据集的结果以及跨市套利结果分别在牛市和熊市两种情况下进行对比分析，发现了基于传统统计方法模型没有发现的可套利组合，说明基于深度学习方法的套利问题研究适用的广泛性。显示 trans_LSTM 融合模型的性能更好，得到更高的收益率，并且符合牛市和熊市回测时间的股票现状，验证了 trans_LSTM 融合模型在股票套利策略中的高效性和适用性，为金融市场中深度学习的进一步发展提供了有价值的参考。

关键词：套利策略 非线性 深度学习 长期依赖

Abstract

Lately, the stock securities market in China has seen robust growth. The slow adoption of arbitrage tactics in the progression of quantitative trading, spanning from theoretical studies to institutional structuring and practical investigation, signifies a pivotal moment in the evolution of China's securities market.The rapid expansion of trading information in financial markets has led to a growing array of elements affecting stock values.As a result, there's been an increased need to analyze technical instruments in the financial sector. The progression in computer technology has enabled the widespread adoption of machine learning techniques, with deep learning, a key element of machine learning, offering substantial benefits in managing nonlinear connections and complex, high-dimensional data.

The Long Short Term Memory Network Model (LSTM) is suitable for dealing with time series related problems due to its excellent recursive properties. However, the LSTM model has its unique structural properties, which cannot fundamentally solve the problem of long-term dependency. The Transformer model is based on self attention mechanism and incorporates parallel attention mechanism, which enables it to capture long-range temporal characteristics and avoid long-term dependencies. Therefore, this article combines the characteristics of LSTM and Transformer models, constructs a trans-LSTM fusion model, and designs

a stock arbitrage strategy based on deep learning, further comparing and analyzing it in detail with LSTM models and statistical cointegration models. The first part of this article provides an overview of the background significance of the research, summarizes the research progress at home and abroad, and provides a concise description of the research content and innovation points of this article. The second part provides a brief introduction to the basic knowledge of arbitrage. The third part introduces the model and strategy for designing arbitrage models based on trans_LSTM. The fourth part is empirical analysis, which involves traversing stock groups and constructing a prediction model based on the trans-LSTM model, LSTM, and cointegration theory to predict the price difference of arbitrage portfolios.

By constructing a complete trans LSTM fusion model and an arbitrage model of LSTM trading strategy, a comparative analysis was conducted on the results of the Science and Technology Innovation Board dataset and cross market arbitrage results in bull and bear markets, respectively. Arbitrage combinations that were not found in traditional statistical methods were found, indicating the widespread applicability of deep learning based arbitrage research. The performance of the trans-LSTM fusion model is shown to be better, resulting in higher returns and consistent with the stock situation during bull and bear market backtesting times. This validates the efficiency and applicability of the trans-LSTM

fusion model in stock arbitrage strategies, providing valuable reference for the further development of deep learning in financial markets.

Keywords： Arbitrage strategy; Nonlinear; Deep learning Long-Term dependencies;

目 录

1 绪论

1.1 研究背景与意义

随着我国金融科技以及其他关键领域的迅速整合，金融科技和金融行业的深度发展正在加快推动金融行业现代化的步伐。在“十四五”计划的实施过程中，金融科技被明确视为推动我国经济向更高水平转型的关键战略领域。在推进“中国式现代化银行”的发展过程中，金融科技发挥了不可或缺的作用，通过数字技术和大数据的深度分析，各大银行逐渐开始推出更加个性化和精确化的金融产品和服务，这不仅增强了风险管理的能力，同时也推动了金融普惠，满足了大众对金融服务多样性的期待。在实际的操作过程中，各大银行都在努力融合人工智能、区块链和云计算等前沿技术，创建数字化运营平台和智能风险管理系统，极大地推动了我国银行业向数字化方向的转型。这一系列措施不仅为提高金融服务水平和促进经济增长奠定了坚实的基础，同时也为金融行业的创新和进步注入了新的活力。金融科技的不断进步不仅仅是服务方式的更新，它更是金融行业变革的引领者，引领整个金融领域步入数字化的新时代。因此，在我国银行业不断向前发展的过程中，金融科技不仅扮演着引领角色，同时也是决定金融未来走向的关键因素之一。

信息技术的不断进步为投资者提供了便捷获取、存储和传播信息的机会，这导致了电子化交易的兴起，进而催生了量化交易的发展。量化交易可以被广泛地理解为一种通过数学模型和计算机技术来实施的交易方式。在目前的情境中，我国普遍采用的量化交易手段有股票多因子策略、期货策略以及高频交易策略等。在 2010 年以前，量化交易依然被视为一个较为小众的行业，随着沪深 300 指数期货的推出，量化交易开始迅速发展。2010 年到 2014 年这几年里，各种量化交易手段都取得了显著的盈利，这促使人们更加重视并投资于此领域，从而使得量化交易的规模迅猛增长。然而，人们对于量化交易存在着一些误解，主要源于过高的期望收益。量化交易作为一种工具，其目的在于提升投资业绩。作为量化交易中的关键策略之一，统计套利交易的起源可以追溯到 20 世纪 80 年代的摩根史丹利，运用数理统计来分析证券价格的历史趋势和差异，并通过设计套利交易策

略来达到盈利的目的。统计套利策略本质来讲是市场中性策略，基于统计学与计量经济学的理论，融合了计算机技术，旨在通过选择合适的投资组合来构建交易配对，从而在相对较低的市场风险中获得稳定的回报。统计套利策略是基于市场的做空机制构建的，选择在被明显低估的资产中设置多头仓位，而在被明显高估的资产中则选择空头仓位。这一市场中性的策略是通过对资产进行差异化的定价，以追求最大的相对收益，从而为投资者提供了一个相对稳定的量化交易路径。

我国的量化交易起步时间相对较晚，但在近几年内，随着金融机制的持续完善，逐步展现出了其活跃性。在股票市场中，大多数投资者主要以中小型为主体，由于在交易中存在非理性行为以及缺乏专业技术、相应技巧和信息获取的难题，导致很难实现较为稳定的收益。例如，2015 年的“股灾”深深影响了投资者的信心，许多人认为是由于股指上的高频交易所导致，因此被视为此次事件发生的元凶。中金所为了应对这种情况，实施了多项策略，例如限制股指的交易频次、增加交易的手续费等，主要是为了降低高频交易的发生。我国证券市场在推进量化交易方面经历了一系列关键时刻，以 2010 年 3 月 31 日为例，我国的证券市场首次推出了融资融券服务。紧接着，沪深 300 股指期货在 2010 年 4 月 16 日正式上市，而上证 50 和中证 500 股指期货则是在 2015 年 4 月 16 日正式上市。期间，A 股市场引入做空策略，这代表了我国证券市场统计套利进一步完善，实现了从理论探索到制度构建再到实际应用的完整流程。这一系列举措成为我国证券市场发展的重要里程碑，为促进量化交易的进一步发展奠定了基础。然而，尽管我国在量化交易方面取得了一些初步成果，但在实践中仍然面临一系列挑战。投资者的非理性行为和信息获取不对称等问题仍然是制约量化交易策略实施的关键因素。因此，为了进一步增强量化交易在我国金融市场的作用和稳健性，我们需要对这些问题进行更为深入的探讨和解决。为了解决信息获取的不平衡问题，要进一步增强监管，确保市场信息能够公正地传达，并采用尖端技术进行更为智能的信息分析。通过融合自然语言处理和机器学习等先进技术，更快速和准确地获取关键数据，进而增强量化交易系统对市场变化的精准把握。

随着计算机技术的持续进步，统计套利方法也在不断地更新和演变，从最初主要依赖时间序列模型的概率统计策略，逐步转变为以数据驱动的机器学习为核心的策略。机器学习在多个领域有着广泛的应用，例如数据挖掘、证券市场分析、

医学判断、无人驾驶和 NLP（自然语言处理）等。其中，深度学习技术在数据分析、图像处理和语音识别等方面展示了卓越的性能。因此，在当前的学术和投资领域，如何采用机器学习算法来制定量化的投资策略将会是备受关注的热门议题之一。基于此背景，将机器学习中的深度学习方法应用到统计套利策略的研究和构建中。首先获取科创板市场以及主板市场 2020 年 6 月 1 日-2023 年 6 月 1 日的日交易数据，将数据进行筛选、归一化等预处理后，计算科创板市场内部价差数据以及主板市场-科创板市场跨市价差数据，构建 trans_LSTM 融合模型对价差数据进行预测，建立套利策略。随后，与传统统计方法进行比较。最后得出相应的结论和建议。

1.2 文献综述

有效市场理论把市场细分为三个不同的级别：弱式有效、半强式有效以及强式有效市场。在弱式有效市场环境下，证券价格不仅全面地反映了历史价格和交易量所隐含的各种信息，而且也限制了投资者通过分析历史价格来获取额外利润的能力，因此必须依靠基础分析来实现投资回报。在半强式有效市场与强势有效市场的背景下，证券的价格完整地展现了所有已经公之于众的数据，这包括交易的价格、交易量、公司的管理状态以及其他已经公开的财务细节。在这样的背景下，依赖基础分析变得不再有效，超额收益的可能性大大降低。但是，在真实的投资场景中，众多的投资者展现出了非理性行为，并且难以迅速地调整他们的交易策略以应对市场的波动。这种情况可能会导致某些信息没有得到充分的体现，从而为套利创造了条件。市场效率理论认为，金融市场反映了所有可获得的信息，并且价格会迅速调整以反映这些信息。尽管如此，套利仍然是一种在非有效市场条件下利用定价错误或不一致性来获取利润的金融交易策略。这一策略通过挖掘市场中的定价错误，利用未被充分反映的信息，试图在市场的非理性行为中寻找套利机会。这突显了市场效率和非理性行为之间的微妙平衡，为理解金融市场的运作提供了有益的观点。吴振翔（2007）针对中国股票市场的弱有效性进行了详细的统计套利检验，并通过假设检验进一步证实了中国股票市场弱有效性的套利检验结果。基于市场效率理论，套利者试图通过分析市场中的统计关系和价格异动来发现市场中的定价错误，金融学强调风险和回报之间的权衡，统计套利通常

涉及较小的价格差异，因此其风险可能相对较低，能在相对较低的风险环境中独立于市场动态并实现较高的回报，是投资机构和学术界关注的焦点。通过对相关文献梳理时发现，统计套利建模主流的方法有距离法、协整法以及多因素模型方法等。

近年来，随着计算机技术的迅猛发展，金融市场逐渐迎来机器学习技术的广泛应用。通过程序化设计，复杂而严谨的数理统计理论在实践中得以全面实现，使得量化投资决策在面对市场环境的变化时更具理性和高效性，这种趋势使得量化投资能够准确捕捉投资机会并有效地控制风险，从而在相对较低的风险水平下实现超额收益。虽然国外在量化投资方面的起步较早，我国由于存在技术含量较高且实施难度较大的问题，仅有部分机构投资者具备使用的技术和资源。然而，近年来，相关理论的深化不断推动实践的发展，取得了显著成果。国内外的研究文献不断涌现，对量化投资进行了广泛而深入的分类梳理。这一系统性的研究助力于更全面地理解量化投资的发展趋势、应用领域和实际效果，为推动该领域的进一步创新提供了坚实的理论基础。综合来看，量化投资作为金融市场中一种创新而高效的投资方式，正不断吸引学术界和实践者的关注。下文将通过对国内外文献的细致梳理，我们有望深入挖掘量化投资的内在机制，为未来的研究和实践提供更加深入的理论指导和实证支持。

1.2.1 国外文献综述

1.基于传统统计方法的统计套利策略

关于统计套利的探讨，目前的研究焦点主要集中在如何将协整理论融入套利模型之中。Engle 与 Granger（1987）首次提出协整理论。协整是指当两个或两个以上的经济变量本身展现出非平稳的特性，但这些变量内部或它们之间的线性组合可能导致序列显示出平稳性，这主要是为了解决在非平稳序列中是否存在长期的稳定均衡关系这一问题。Burgess（1999）利用金融时报 100 指数成分股数据，以协整理论为研究基础，构造了相对价值关系的统计套利模型，当组合价格与长期趋势均值偏离时，此时可以通过套利模型获利。Vidyamurthy（2004）通过将协整理论与统计套利结合起来，并利用股票价格数据中内部和外部的协整关系，通过协整系数和均值来建立股票价格的线性关系，从而制定出一个合理的配对交易

策略。Rudy 等（2010）在高频股票数据中运用成对交易统计套利技术，通过与日收盘价标准采样频率对比得出盈利潜力，观测频率区间为 5 分钟区间至各交易日收市后所记录价格，结果发现，当多元配对交易策略运用到高频数据上时，配对数据内和数据间协整程度较高，所得总收益越大。Fan（2017）对豆油和棕榈油期货的统计套利机会进行了深入研究，研究发现，在传统的协整模型中，残差估计并不是稳定的，但当采用贝叶斯方法进行参数估计时，得到的残差序列是稳定的，从而获得了更为优越的统计套利效果。除了协整方法之外，另一些方法受到了学者们的重视。Gatev（2006）与 Do（2010）都通过距离法来研究统计套利问题，Gatev 用美国股市 1962-1997 年的日数据根据归一化价格最小距离匹配成一对建模来进行统计套利策略研究，最终得到了组合瓶蕨年华超额收益达到 12%的结果。而 Do 在此基础上进一步研究，将时间延长 7 年发现此时收益是下降的，进一步原因分析中得出主要在于距离法并没有考虑到长期能否发展的情况。Elliott（2005）提出了均值回复高斯马尔可夫链模型研究配对交易的基本框架。Bogomolov（2013）使用非参数交易的方法构建了一个基于 O-U 过程的策略，对美国和澳大利亚证券交易所的每日市场数据进行的测试显示，平均超额收益率为每月 1.4-3.6%，年化夏普比率为 1.5-3.4。

2.基于机器学习的套利策略

Dunis 等（2006）基于人工神经网络模型，建立了价差的配对交易策略；Thomaidis（2006）提出基于 NN-GARCH 模型的套利策略；Fischer 与 Krauss（2017）利用长短期记忆网络对标普 500 指标的成分股进行了详尽的实证研究，结果显示了 0.46%的盈利率以及 5.8 的夏普比率的实际表现。Dunis 等（2008）使用协整和三种机器学习方分别对法国石油期货数据采用交易过滤器对模型进行优化，结果显示带有跨线性滤波器的多层感知器神经网络效果最好。Huang 等（2015）、Lin 与 Cao（2008）利用遗传算法对筛选出的股票对进行分析，得到高收益回报的结论；Kim 等（2019）推出了一个利用关系数据的分层注意网络股票预测，旨在预估个股价格及市场指数的波动。Montana 与 Parrella（2009）利用支持向量回归对近 7 年 S&P500ETF 的历史数据进行套利研究，证明了这种交易策略可以获得回报；Gu 与 Bryan（2020）他们采用了机器学习技术来研究资产的风险溢价，并进行了实证对比分析，利用机器学习方法可以更好地帮助我们实证地理解资产

价格。Mulvey 等（2020）通过采用 FNN（前馈神经网络技术），利用均值回复对部分符合条件的投资组合进行了深入的实证分析，实验结果证实，深度学习在捕获时间序列套利机会方面表现出了显著的有效性。Dunis 等（2015）使用多层感知器神经网络并结合使用高阶神经网络，与遗传编程算法进行对比实验，构建玉米/乙醇压榨价差的配对交易策略；Kim（2022）构建一种名为 HDRL-Trader 的混合深度强化学习模型，旨在解决当前基于强化学习研究中现有模型存在的问题。该模型具备用于识别交易行为和用于确定止损的界限的独立学习网络，结论显示在平均回报率方面实现了一定程度的提升。Qing 等（2021）使用订单簿背后所映射的市场价格的高频数据类型，通过构建新型的订单流不平衡指标讨论套利策略，对比实验的结果显示，新的指标在性能方面相较传统指标而言展现出了更出色的表现。Qiu（2020）利用小波变换（WT）对股票价格的原始数据进行了初步处理，然后采用基于注意力机制的长短时记忆网络模型对其进行了预测，与标准的长短时记忆网络模型、门控循环单元模型进行了对比，研究结果显示该模型具有更高的预测准确性。Jorge 等（2021）构建了一个统一的统计套利理论框架，并引进了深度学习的解决策略。通过从大规模面板数据中识别共性和时间序列模式，证实了将深度学习等高级技术应用于统计套利具有极大的应用潜力。

1.2.2 国内文献综述

1.基于传统统计方法的统计套利策略

丁秀玲与华仁海（2007）基于协整理论，我们对大连商品交易所的大豆和豆粕期货价格的变化趋势以及套利交易行为进行了深入探讨，研究结果发现这种套利交易的盈利能力并不明显。于玮婷（2011）通过协整技术，我们选择了 90 只标有我国融资融券的股票，并采用成对交易的统计套利策略，实验数据证实了优于不采用策略方法时得到的收益，证明了该模型在我国股票市场上的适用性。赵胜民与闫红蕾（2015）基于转移模型对我国主板市场中的可融资融券的股票数据进行价差收敛性检验，得到相同行业板块的股票在进行统计套利时组合而成的股票对或许会存在相对较高风险的结论。胡伦超等（2016）结合距离方法和协整方法，在针对上证 50 成分股的实证研究中，我们构建了一种两阶段配对交易策略，并发现通过这种方法筛选出的组合能明显地获得更高的回报。葛翔宇等（2012）

在 Balke 研究的基础上提出了基于门限向量误差修正模型的检验，建立关于 UK100 和 GER30 的套利模型，检验了不同市场上的同质或者相似商品的价格存在长期均衡关系。袁晨与傅强（2017）通过建立 DCC-MVGARCH 模型，对沪深 300 和上证 50 的期货与现货之间的关联性以及套期保值的效果进行了深入研究。张戡等（2012）在对主板市场的股票进行聚类分析后，基于协整理论研究发现二者结合可以更好地识别高相关度股票价格序列内部的规律变化，从而筛选出相关性较高且具有协整特性的股票组合，这些组合将明显带来更高的收益。安云博（2013）基于协整理论的统计套利模型对我国证券市场上能够融资融券的 16 支银行股进行了样本内检验，结果显示最终的收益并没有长期持有时的收益高。张河生与闻岳春（2013）选取的 IF1209 和 IF1212 做配对交易通过实证分析通过对参数调整获得最优交易区间和止损区间，使交易策略达到更好的效果。李世伟（2011）对协整理论套利方法进行改进，得到了更好的套利效果。华仁海与仲伟俊（2002）采用协整理论，并结合 GS 模型与误差修正模型，对上海期货交易所的金属铜和铝的价格发现功能进行了深入的实证研究。Li 等（2014）研究显示，采用协整方法的配对交易策略在主板市场和香港 H 股市场的平均年化超额收益大约是 17.6%。欧阳红兵和李进（2015）将最优阈值选择问题转化为利润最大化的问题，针对我过 A+H 股股价数据进行实证分析，结果表明通过对参数的优化可以获得更大的套利空间。朱丽蓉等（2015）以棉花期货作为主要研究内容，并采用了经过优化的跨期套利交易方法，经过实证分析，证明了该模型在中国商品期货市场上的适用性。邢知与郝继升（2018）对沪深 300 股指期货与现货的关系以单位风险下的收益作为优化目标，构建了协整-GARCH 模型的统计套利策略，研究结果显示，通过不断优化模型参数，可以得到更优的建仓阈值和止损阈值。蔡燕等（2012）基于协整理论，并在 Elliott 的理论框架下，引入了价差服从 O-U 过程来研究沪深 300 股指期货和上证 180ETF 的交易配对。研究结果显示，利用价差服从 O-U 过程模拟得到的数据明显优于 Elliott 的分析结果。同样的，黄晓薇等（2015）提出了一种基于 O-U 过程的配对套利策略。研究结果表明，从价格对信息的敏感性、价格平衡的内部机制、市场风险和交易成本四个方面，该模型能够提高市场效率，并且与传统的协整套利模型相比，成本更低、风险更小、收益更高。何至静等（2020）当 GARCH 模型的表达式满足 $\alpha+\beta<1$ 的限制条件时，

可以得到一个随时间变化而变化的价差序列动态标准差，并以此作为交易的信号。相较于传统的参数固定上下浮动 1.5 倍的标准差的情况而言，此方法在夏普比率上的表现显得更为出色。张鹰（2023）以工商银行和建设银行实际数据为基础，基于统计套利的理论基础、定义并结合相对应的计量模型-GARCH 进行协整分析。

2.基于机器学习的套利策略

刘阳等（2016）将神经网络和动态 GARCH 模型结合，考察不同商品内部及之间的组合套利机会，得到的实验结果显示基于神经网络和动态 GARCH 模型的结合模型的套利策略在发现在参数调整下盈利能力显著提高。周亮等（2022）使用了八个机器学习模型及自回归移动平均模型（ARIMA）对不同频率信号进行分析，研究结果表明所有模型均能取得较高的套利收益，将非线性模型和线性模型融合使用能够改善模型的风险控制能力。李斌等（2019）利用 1997-2018 年间的主板市场数据进行了深入的实证研究，研究结果揭示了机器学习算法可以实现超额的收益，并且其投资策略相较于传统的线性算法和所有单一因子都展现出更出色的投资表现。Xu（2020）利用深度学习技术对金融的时间序列数据进行了深入探讨，并设计了一个与卷积神经网络技术融合的混合模型。Firuz（2020）使用了多层感知机、卷积神经网络和长短期记忆网络这三种神经网络模型，对四家美国上市公司 10 年期的日股价数据进行了实验分析，成功预测了股票价格的重大波动，实验结果不仅精度极高，而且明显优于之前的研究成果。陈卫华与徐国祥（2018）利用 LSTM 神经网络模型，以东方财富网股吧日发帖数为研究特征，对沪深 300 指数的已实现波动率进行了预测分析，LSTM 神经网络模型在预测效果上明显优于其他模型。邓晓卫与章铖斌（2019）利用构建 BP 神经网络与 LSTM 神经网络的混合模型，对中国银行股票进行实证分析，结果表明混合神经网络模型的套利策略的结果更为精准。杨青与王晨蔚（2019）通过建立 LSTM 模型来研究全球 30 个股票指数在不同时段的预测，研究结果揭示了 LSTM 神经网络模型在股票指数预测方面具有更高的准确性和适用性，同时其波动性也相对较低。杨云飞等（2010）采用基于经验模式分解（EMD）和支持向量机（SVMs）的非线性组合预测技术来预测原油价格，预测结果表明，这种混合模型具有很高的预测准确性。胡文伟等（2017）将 Sarsa 强化学习算法和 ε-greedy 策略结合成自适应

配对模型进一步对传统统计方法的策略进行改进，模型参数优化设定为自适应动态优化参数法，实证结果表明，使用自适应动态优化参数法所得到的绩效更优。

1.2.3 文献述评

通过上述文献梳理，我们可以发现，随着量化交易的广泛采用，该策略的盈利潜力受到限制，金融市场中的套利空间逐渐减小，导致收益率显著下降，在这种情况下，实施该策略在交易中难以达到理想的效果。另外多数统计套利策略都是建立在协整理论基础之上，无法得到稳定回报。这是由于协整理论在实施的过程中需要满足许多假设条件，如协整理论只讨论线性问题，可能造成数据组的缺失。因此，基于协整理论的统计套利策略研究可能导致套利范围的减少，从而给投资者带来经济损失。由于金融市场具有低噪声和高动态特性，这导致策略在交易行为上显得僵化，从而在长期内难以保持稳定的正向收益。另外，传统统计套利策略常常使用主观经验或固定常数来确定开平仓阈值等参数，然而，研究表明这种基于主观经验或固定常数的参数确定方法存在许多限制。在实际操作中，金融资产的价格时间序列数据呈现出明显的异方差特性，因此，在套利交易策略的设计过程中，交易逻辑需要进行灵活的调整以适应市场环境的不断变化。因此，众多学者已经提出了多种优化建议，包括采用 GARCH 模型和 O-U 过程等方法。然而，这些改进方案往往伴随着新的假设和经验参数的引入，限制了其在变化莫测的金融市场环境中的适应性，因此对策略的动态调整和自我修正仍然面临挑战。伴随机器算法持续的优化，众多研究人员开始逐步将机器学习的方法融入到套利策略之中。实证研究显示，随着机器学习的快速更新与发展，机器学习优越地解决金融数据非线性、高度噪声和虚假相关性等问题的能力，使得更多的机器学习方法在统计套利策略中具有广泛适用性，同时也可能带来新的方法与思路。

长短期记忆网络模型（LSTM）在金融领域的研究中表现卓越，由于金融数据通常具有时序依赖性，LSTM 以其出色的依赖关系处理能力，在揭示金融市场中的显性和隐性信息方面发挥着关键作用，通过更为深入和全面的方法来挖掘时序数据中的有价值的信息，为投资者提供了更为综合和多元的数据分析视角，从而更有助于做出更加明智和准确的决定。随着 Transformer 方法的不断发展，自注意力机制的引入成功学习了复杂的模式和动态变化，在自然语言处理和股价预

测等领域取得显著成绩。为充分发挥两种模型的优势构建 trans_LSTM 融合模型，并将其应用于统计套利策略的分析中，采用收益率水平作为衡量套利策略优化程度的准则，有助于我们更深刻地洞察市场信息的传播和交易行为模式。通过结合 LSTM 和 Transformer 的预测能力，这一方法为深度学习在金融领域的进一步发展提供了新的途径。trans_LSTM 融合模型的应用不仅为投资者提供更多参考，帮助其实现超额收益和风险分散的目标，同时成为了解市场有效性的一种途径。在控制金融风险方面，这种方法具有重要的意义。通过深度学习方法，特别是结合不同模型的优势，我们可以更准确地把握市场的变化，从而更有效地规避金融风险，为投资者提供更加稳健和可信赖的决策依据。

1.3 研究内容和研究框架

1.3.1 研究内容

本文主要聚焦于套利的概念、种类和特性，探讨了统计套利的发展进程，并深入研究了长短期记忆模型（LSTM）的算法结构以及 Transformer 模型的结构。在这个基础上，文章构建了一种基于 trans_LSTM 融合模型的统计套利策略，并通过实证分析在我国科创板市场不同数据集上的应用以及主板市场与科创板市场上的跨市套利问题的研究。研究的核心是将深度学习方法引入统计套利策略的领域，借助 trans_LSTM 融合模型对价差进行预测。该模型结合了 LSTM 和 Transformer 的优势，以更好地捕捉时间序列中的关键特征和长期依赖关系。通过这一方法，我们设计了一套有效的套利策略。实证分析阶段，对比股票市场牛市和熊市两种股票不同情况，以我国科创板市场不同数据集下的情况为例，将深度学习方法应用于统计套利策略的研究中。为了验证其效果，我们将深度学习方法的套利策略结果与传统的统计分析方法进行了比较。最终，进一步通过分析研究主板市场和科创板市场之间的跨市套利，对比实证结果，我们得出了关于深度学习方法在统计套利中的应用效果的结论，并对这一研究的意义和未来工作方向进行了探讨。本文的研究为深度学习在金融套利领域的应用提供了一种新的思路，并为相关领域的研究和实践提供了有益的启示。

1.3.2 研究框架

本文的研究框架如下：

第一章：绪论。这部分主要阐述了这篇论文的研究背景及其研究意义。其次详细描述了套利背后的市场驱动因素，以及国内外学者对套利议题的当前研究进展。最后描述了研究的核心内容，并阐述了论文中的创新之处。

第二章：套利的基本知识。这一部分介绍了套利的定义与类型以及套利策略的发展，包括主流方法统计套利策略以及当前机器学习方法套利策略。

第三章：深度学习套利模型设计。这部分介绍了长短时记忆模型以及Transformer模型的结构，随后通过二者模型结构优势进而提出了构建tran_LSTM融合模型结构，并介绍了融合模型各层的具体细节。随后介绍了本文使用的评价指标，包括预测指标和策略指标两种。

第四章：基于 trans_LSTM 的套利策略实证分析。这部分首先介绍了本论文使用的数据及变量，并且对数据进行简单的清洗和归一化等预处理的基础工作，以便更好的符合模型训练要求。随后对价差数据分别带入不同模型进行价差的预测，选择股票对在牛市和熊市两个不同时间段进行回测，继续下一步的分析。通过对比 trans_LSTM、LSTM 套利策略的收益率指标的结果，证明本文所提出的混合模型的效果和性能是最佳的，并且发现了基于协整理论的统计套利策略未能发现的套利空间。进一步研究模型的适用性，拓展模型使用范围，分析研究了主板市场和科创板市场的跨市套利，同样得到了相同的结论。

第五章：结论与展望。该部分总结以上主要研究内容及实证结果，并分析研究过程中存在的问题，同时对今后研究方向做出展望并提出进一步完善方向。

1.4 创新点

本文创新地将 Transformer 与 LSTM 模型结合起来，充分发挥 LSTM 记忆性强的特点，提出了 trans_LSTM，并对模型预测效果的改善与否进行了观察。文章通过对价差数据预测，采用套利策略中的交易配对的思想，来对科创板市场股票数据进行问题的分析，并且加入跨市套利。研究方法上应用基于深度学习的非线性模型进行套利策略研究。关于套利问题的研究，在内容上前人的研究多是针

对期货市场进行的策略问题的研究，在方法上多利用的是基于协整理论的协整模型。Transformer 模型在 CV 和 NLP 等领域应用非常广泛，表现显著的性能提升，而金融领域分析中的应用极少。Transformer 模型可以在整个序列上进行并行计算，而不需要按顺序处理每个位置，由于自注意力机制的特性，每个位置的表示可以同时考虑整个序列的信息，在处理大规模数据时具有较高的效率。本文试图将 Transformer 应用到金融领域中来。LSTM 是一种深度学习模型，尤其适用于序列数据的处理和加工，它所具有非线性激活函数和复杂的内部结构，可以捕捉序列数据中的非线性关系和模式。相比于传统的线性模型，如线性回归，LSTM 具有更强大的建模能力，可以更好地适应具有非线性结构的数据。这使得 LSTM 在处理时间序列数据、自然语言处理等领域中非常成功，因为这些领域的数据通常包含复杂的非线性关系。因此，在不损失其数据特点的情况下，LSTM 模型在金融领域可以更好的发挥其优势。

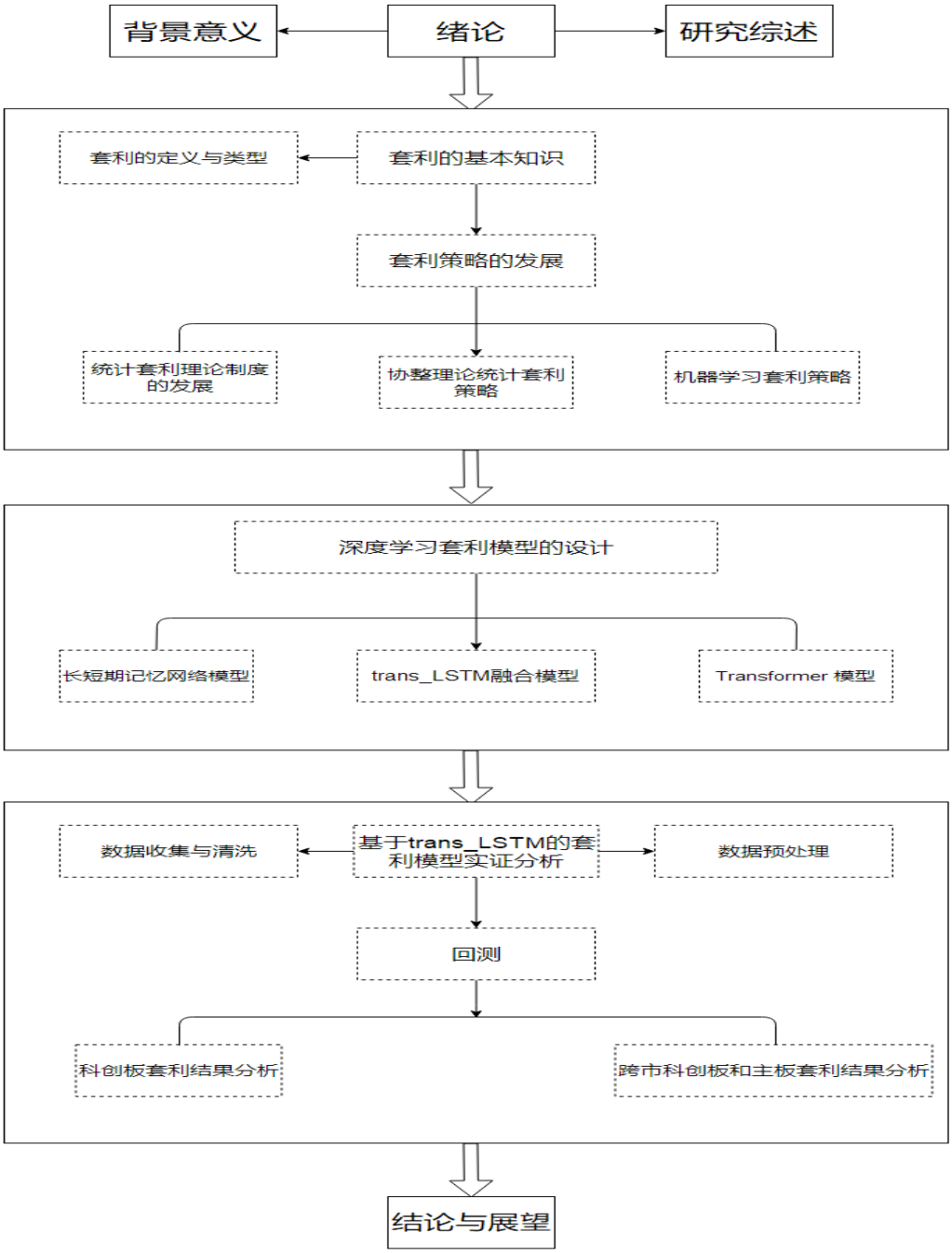

背景意义
绪论
研究综述
套利的定义与类型
套利的基本知识
套利策略的发展
统计套利理论制度的发展
协整理论统计套利策略
机器学习套利策略
深度学习套利模型的设计
长短期记忆网络模型
trans_LSTM融合模型
Transformer 模型
数据收集与清洗
基于trans_LSTM的套利模型实证分析
数据预处理
回测
科创板套利结果分析
跨市科创板和主板套利结果分析
结论与展望

图 1.1 技术路线

2 套利的基本知识

2.1 套利的定义与类型

2.1.1 套利的定义

套利是指在一个或多个市场中，当某一实物资产或金融资产有两种不同的价格时，通过以更低的价格购买并以更高的价格出售，从而达到零风险收益的交易方式。套利策略的核心目标是修正市场价格或收益率的不正常状况，并利用这些不正常的数据来实现盈利。在某些特殊情况下，同一种产品在不同的市场上可能会有明显的价格差距，而套利策略则是通过以较低的价格购买并以较高的价格出售，从而使价格达到一个平衡状态。套利的典型做法是在特定的市场或金融工具上设立头寸，随后在其他市场或金融工具上设立与之前头寸相平衡的头寸。当价格恢复到一个平衡状态后，所有的头寸都可以被结算，从而达到盈利的目的。套利策略的核心思想是利用各种市场或不同类型的相似或同类金融产品之间的价格差距，通过购买价格较低的合约并同时出售价格较高的合约来实现盈利。在进行套利操作时，交易者更多地关心合约间的价格相对关系，而不是价格的绝对水平。在最理想的状况下，套利操作应当是完全无风险的。

套利交易作为一种金融交易手段，其核心理念是通过利用不同市场中相关证券的价格差异，在购买和出售相应证券的过程中获得利差收益。起初，套利交易主要集中在那些风险极低或完全无风险的交易手段上。但是，随着市场交易模式变得越来越多样化，一些事件性交易策略和短线交易策略也被归类为套利交易，这使得对套利交易模式的统一定义变得更加困难。套利交易的目标在于通过充分利用不同市场、资产或合同之间的价格差异，实现较低风险、相对稳定的回报。这种策略的关键在于同时进行买卖操作，以确保在市场波动中能够有效地捕捉利润机会。由于套利交易的本质是基于市场不完善性和价格差异的存在，它通常具备一些特征，如低风险、即时性、市场不完善性以及合规性监管等。套利交易的优势在于其能够提供一种相对较为稳健的投资方式，尽可能规避市场波动带来的风险。然而，这也要求套利交易者对市场的敏感性和变化有着深刻的理解，并能够快速做出决策以把握时机。此外，随着市场的不断演变，套利交易的定义也随

之变得更加多元化，需要灵活应对各种交易策略和市场情境。

在过去，套利是机警交易员常用的交易手段，但随着技术的不断进步，现在已经演变为一种依赖于复杂计算机程序的交易策略，让交易者有机会从不同市场中的同一证券的细微价差中获得盈利，从而提升了套利操作的效率和准确性。

2.1.2 套利的类型

在金融经济领域，套利行为的分类标准各不相同，通常是基于时间、地点和商品的特性来进行的。它的主要分类可以分为三个大类：

1.跨期套利

在金融市场中，交易的商品在不同的时间段会出现不同程度的价格波动，这为套利者创造了特殊的机会。跨期套利是一种利用这种时间差异的策略，套利者通过及时购买低价商品和抛售高价商品来获取价格差异带来的收益，这种行为的基本原理是在商品价格非正常波动的情况下，充分利用市场的不同期限或时间点的价格差异。跨期套利的策略涵盖了迅速购买低价商品和及时出售高价商品，目的是为了充分利用价格之间的差异，这个策略的关键是要捕获商品价格的不稳定波动，以便在市场的不稳定中获得经济利益。除此之外，跨期套利也表现为金融机构基于其融资优势进行的高利率贷款和低利率融资活动。通过这种方式，企业可以在借贷和融资的过程中巧妙地利用利率差异，实现更为优化的融资结构，从而提升经济效益。

总体而言，跨期套利是一种充分利用不同时间段内商品价格波动的策略，其操作涉及购买低价商品、售出高价商品，并且在金融领域中还包括对利率差异的灵活利用。这种套利行为旨在借助市场的非正常波动，实现经济利润的最大化。

2.跨市套利

随着商品在地理位置上的变动，其价格也可能出现显著的波动。在不同的地理区域，外部环境因素如金融市场所接受的政策和法规、市场机制等都存在显著差异，这些因素直接导致了商品价格的多样性和不同程度的波动趋势，套利者借助这种情况，进行了跨市场的套利活动，并在各个地区的价格差距中实现了盈利。跨市场套利的策略主要涉及将商品从某一特定地区转移到需求量较

大的区域，并随后以更高的价格进行销售。套利者通过这种方式，充分利用了不同地区之间商品价格的差异，实现了从价格波动中获取经济利益的目标。这种跨市套利的行为基于空间位置的不同，使得商品在不同地区的供需状况和市场环境存在显著差异。套利者通过迅速反应和巧妙操作，能够在这一差异中找到机会，实现商品从低价地区到高价地区的转移和价差收益的获取。

总体而言，跨市套利是一种基于地区之间商品价格差异的策略，套利者通过将商品从一个地区带到需求量大的地方高价售出，巧妙地利用了地区之间的市场不同，实现了获利的目标。

3.跨商品套利

跨商品套利被视为一种更加独特的套利活动，要求更严格的前置条件和对产品的精确筛选。在众多的商品当中，它们之间可能有某种深层次的联系。套利者可以通过深度探索商品间的互动关系，巧妙地运用这些联系来达到盈利的目的。这种套利行为要求套利者对多种商品之间的潜在联系有深入的理解。举例而言，考虑到灯泡和蜡烛这两种商品，在停电时，蜡烛的销量可能会急剧增加。套利者通过对这些商品之间的关系有所了解，便可以在特定情境中实施跨商品套利策略。跨商品套利的成功实施需要套利者具备对市场和商品的深刻洞察，能够精准把握商品之间的内在联系和相互影响。通过对这些联系的细致分析，套利者能够灵活地选择并操控商品的交易，以在市场波动中获取差异化的收益。

从宏观角度看，跨商品的套利行为需要更高的前置条件和更深入的理解。要求套利者具备更加扎实的数学基础以及良好的金融背景知识。套利者能够通过深入理解商品间的互动关系，在特定的情境中巧妙地运用商品间的相互联系，从而实现更加精细的盈利目标。

2.2 套利策略的发展

2.2.1 统计套利理论制度的发展

统计套利作为一种金融套利手段，其核心思想是运用统计学和数学模型来探究不同资产间的价格动态或统计失衡，进而通过构建投资组合以实现稳健的盈利。

在美国，统计套利被广泛用于对冲基金，对冲交易等市场中。这一投资策略采纳了一种独特的投资策略，它依赖于不同投资品种间的相互关联性来进行操作，尤其是在两个投资品种的价格差异扩大至正常范围时。在执行统计套利的过程中，投资者会在价格差距扩大到特定程度时，通过同时建立多个空头寸来进行多空仓位的操作。当价格差异恢复到正常范围后，投资者会进行双向平仓操作，这样可以避免系统性的风险，并从两者之间的短期价格差距中获得利润。这一策略的核心目的是在资产组合中识别出不正常的定价模式，并在实际价格与理论价格出现偏差时设立头寸，等待其回归到理论价格后执行平仓操作。统计套利的主要过程包括发现资产组合的异常定价关系、建立头寸以及在价格回归到理论水平时进行平仓。通过精确地识别和利用不同资产之间的关系，统计套利策略旨在实现相对稳定的投资收益，同时通过灵活的多空操作方式规避潜在的系统性风险。

统计套利是一种常见的套利方法。从宏观角度来看，其在交易操作中具备如下流程：通过利用历史数据进行深入的统计分析，进而发现其中存在的规律性和发展趋势，根据所发现的规律与趋势构建投资组合对以及适合的模型，最后将这些应用于实际公开市场操作中，利用统计规律来探寻市场中可能存在的套利机会，并在合适的时间点上进行买卖操作从而实现投资收益。统计套利的核心思想是依赖历史数据中的规律，通过对市场的深度分析和理解，找到潜在的价格差异，并通过及时的套利操作获得利润。这种策略注重对市场趋势和统计规律的准确把握，以期在市场波动中稳健运作，实现资产的增值。统计套利策略是一种中性的交易方式，其交易信号是基于一定的规则所产生，另外策略所获得的收益是通过统计手段所得到的（Avellaneda 与 Lee, 2010）。

随着中国股票市场的融资融券制度日益成熟和股指期货的加入，中国的做空策略逐步完善，为统计套利策略创造了更加有利的条件。这使得统计套利策略在中国的股票市场上逐步显现其重要性。其市场中性和基于规则的交易信号使得统计套利策略在不同市场环境中能够相对独立地运作，更加适应快速波动的市场情况。因此，这一策略在我国市场的发展可能会受益于市场结构的不断改善和制度环境的进一步完善。学者们对统计套利理论的不断深化研究，统计套利的理论框架也逐渐走向成熟和完善。Moegan Stanley 把统计套利看作是一种模型过程，在不受经济因素影响的前提下，通过量化的方式来构建资产组合，其核心思想是一

种所谓的“配对交易”。在配对股票受到短期因素冲击时，当配对股票价差暂时偏离其长期均衡关系时，统计套利策略通过买入低估的股票和卖出高估的股票来利用这一价差。

S.Hogan，R.Jarrow 和 M.Warachka（2004）对于统计套利给出了更加精确的数学定义的公式，并着重强调了统计套利作为一种具有零初始成本和自我融资特性的交易手段。当我们使用$H(t)$来表示某一 t 时刻的累计收益时，无风险收益率折现的现值$h(t)$，满足如下条件:

（1）$H(t)=0$，表示初始成本为零。

（2）$\lim_{t\to\infty} E\big(h(t)\big)>0$，表示组合的收益平均值的极限大于零。

（3）$\lim_{t\to\infty} P(h(t)<0)=0$，表示组合亏损的概率趋向于零。

（4）若$\forall_t<\infty$，$P(h(t)<0)>0$，则 $\lim_{t\to\infty}\frac{var(h(t))}{t}=0$，这意味着，在一个有限的时间段内，如果损失的可能性是正的，那么收益的方差将会相对于时间收敛到零。

通过上述统计套利所满足的四个数学公式可以看到，统计套利与无风险套利实际上是存在显著的差异。与无风险套利相比，统计套利涉及到买入和卖出的双向操作，其固有的风险则会随着时间的推移逐步降低，最终可能接近零。在实践中，为了执行这一策略，投入相应的资本是不可避免的，包括但不限于交易成本和保证金。这与无风险套利的零投资特性形成了鲜明对比。尽管统计套利的投资可能会伴随一定的费用，但这也为套利者提供了机会通过巧妙的交易策略获得收益。值得注意的是，统计套利的收益在很大程度上是独立于市场行情的。与一些传统投资策略相比，统计套利对市场波动的敏感性较小，使得其风险在一定程度上可以被控制。这表明统计套利策略不仅仅依赖于市场的整体走势，而更注重证券价格的历史统计规律，从而在一定程度上减轻了市场波动对套利行为的影响。

2.2.2 协整理论统计套利策略

协整套利策略是建立在协整关系理论基础上的一种金融套利策略。协整套利的理论思想是通过发现和利用市场上存在的协整关系，建立相应的交易策略，以期望在价格偏离协整关系时获得套利机会。在策略构建的过程中，我们使用协整

测试来鉴定那些具有协整特性的股票，并在交易过程中采用简洁的方法来产生交易信号，这些信号大部分是基于 GGR 闽值规则的。协整法采用计量经济学手段来确定配对组合，并首先利用协整检验来识别那些在价格或收益之间具有长期稳定性的目标资产。这批资产的协整方程残差展示了稳定的均值恢复特性，因此具备一定程度的预测能力。基于这一发现，该策略将确定开平仓的阈值，而这个阈值是基于残差的波动特性来确定的。在残差触发开仓阈值的情况下，策略会通过同时购买被低估的资产并出售被高估的资产，然后根据这些资产的协整系数来确定投资组合的比重。这种方式形成了一个综合性的资产组合。在这一复杂的资产组合里，当残差达到平仓的阈值，策略会选择反向平仓，从而解决现有的头寸问题。整个过程中，套利者利用协整关系和残差的统计性质，寻找标的资产之间的价差变化，并基于这些变化进行有预测性的交易。这种套利策略的关键在于建立有效的协整关系模型，准确地设定开平仓阈值，并根据协整系数确定合适的投资组合比重。通过这一系统性的方法，套利者能够在价格波动的背后找到有潜在收益的机会，实施精准的交易。

这一策略的发展源于对市场协整性质的深入研究，主要基于经济学中均衡和长期关系的概念，是经济学和计量经济学中的一个重要概念。协整关系理论的核心观点是，虽然两个或更多的时间序列可能表现为非平稳状态，但它们之间有可能存在一个稳定的线性组合，而这个组合本质上是平稳的。协整关系揭示了这些时间序列间有一种平衡的联系，使得它们在长期内能够共同演变，为统计套利提供了机会。套利者通过捕捉资产价格的长期均衡状态，在价格偏离均衡时执行交易，实现套利收益，因此协整关系为套利提供了理论支持。依据 CAPM 模型的分析，一个市场内的每一只股票都与市场的基准指数有一定程度的相关性，因此，在任何两只股票之间也存在相应的相关性。当两只或更多的股票的股价展现出长时间的线性稳定关系时，我们可以推断它们之间有协整的联系。当股票价格在短时间内偏离了这一均衡状态，就会有一个修正机制将这种偏离调整到一个合适的范围内。

协整套利策略得益于计量经济学和量化金融领域的研究成果，尤其是对协整关系检验和协整向量估计方法的不断改进。随着计算技术的提升和数据的广泛可得，套利者能够更有效地发现和利用协整关系，推动了该策略的发展。然而，协

整套利策略在面对市场变化、模型拟合不足和交易成本等方面仍然面临挑战，特别是在处理非线性、非平稳和高频数据方面存在一定的局限性。

2.2.3 机器学习套利策略

随着科技的飞速发展，机器学习在金融领域的应用已经成为一种不可忽视的趋势，尤其是在套利方面。这一趋势的崛起为金融从业者提供了新的工具和技术，以更加智能和高效地进行投资和交易。金融市场充满了海量的复杂数据，包括市场价格、交易量、财务报告等，通常呈现出时间序列的结构，具有高度动态性和复杂性。价格波动、趋势、季节性和突发事件等因素都在数据中反映出来。随着统计套利策略的广泛应用，其盈利空间不断受限，与此同时，随着机器算法的持续发展，更多的学者开始关注机器学习，在量化交易方面，机器学习的应用尤为显著。通过建立预测模型，识别价格波动、趋势和其他市场信号，量化交易策略得以更为精准和高效地执行。这种自动化的交易方式使得投资者能够利用大数据和先进算法迅速做出决策，降低人为因素的影响。将机器学习引入统计套利策略，通过对金融数据的深度分析，能够揭露潜藏其中的各种模式与发展趋势，为决策提供有力支持。数据驱动的决策成为金融从业者的主要策略之一，帮助他们更好地理解市场动态。通过对国内外文献梳理大量实证研究表明，机器学习在统计套利策略中具有良好的适用性，与协整策略相比，机器学习策略在套利机会发现、风险控制和适应市场变化等方面具有一些独特的优势，更好的适配金融数据的特性。机器学习模型能够更全面地考虑多模态数据，提高对市场全貌的解读，其自动化特征提取减轻了对人工特征工程的依赖，使得模型更为全面、准确地表达金融市场中的价差情况。相对于传统协整策略，机器学习策略更具灵活性和适应性，能够更好地适应市场动态。

利用机器学习模型构建套利策略首先，金融市场的价差可能源于不同资产之间的基本面差异或市场交易中的瞬时波动。机器学习模型能够捕捉这些价差的动态特性，而且对于不同资产之间的相关性变化要有灵活的适应性。通过机器学习模型，可以有效地改善传统模型中使用固定标准差作为交易阈值参数 σ 带来的限制，可以捕捉金融数据的非线性、动态的特征，解决数据“扎堆”的现象。其次，对于价差套利策略，时间敏感性和实时性是至关重要的。市场中的价差可能是瞬

时的，而机器学习模型需要能够在短时间内作出决策并执行交易。另一方面，机器学习可以更好的应对金融数据的非结构性特征。比如可以利用情感分析和舆论数据的整合为价差套利提供额外的信号源，尤其是在市场出现非理性波动时。或者通过运用自然语言处理技术，模型可以更好地理解市场参与者的情绪和看法，从而更准确地预测未来的价差变化。另外使用机器学习模型分析研究统计套利问题时，风险管理也是价差套利策略设计中的一个重要方面。机器学习模型能够评估交易的潜在风险，并在制定交易决策时考虑到这些风险。比如考虑到不同资产的波动性、相关性和市场流动性的动态监测，以确保在套利过程中能够有效管理风险。

机器学习利用价差的套利策略设计考虑到了金融数据特性、市场环境和不同资产之间的动态关系。通过整合时间序列分析、高频交易模型等多种技术手段，以及灵活的风险管理方法，设计出适应市场变化、具备实时性的套利策略将更有可能在不同市场环境中取得成功。

3 深度学习套利模型设计

3.1 长短期记忆网络（LSTM）

在深度学习的领域，特别是在循环神经网络（RNN）中，“长时间的依赖”是一个普遍存在的问题。长期依赖性产生的根本原因在于，当神经网络中的节点经历了多个计算阶段后，之前相对较长的时间片段的特性已经被掩盖。随着数据时间片段的逐渐增多，RNN 失去了获取如此远距离信息的能力。

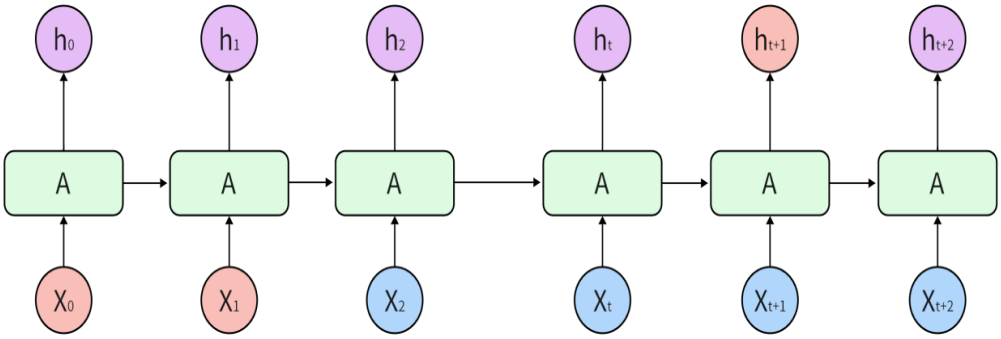


图 3.1 RNN 的长期依赖问题

梯度消失和梯度爆炸是主要影响 RNN 模型训练的因素之一，这些问题主要源于 RNN 中权值矩阵的循环相乘，多次组合相同的函数可能会导致极端的非线性行为，每一个时间步骤都采用了一致的权重矩阵，这使得梯度消失和梯度爆炸的问题变得更加突出。

长短期记忆网络（Long Short-Term Memory，简称 LSTM）这是一个深度学习的神经网络框架，能够处理长短期的信息，通常用于处理和预测与时间序列相关的数据，如语音识别、文本生成、股票价格预测、自然语言处理等任务。LSTM 最初是在 1997 年由 Hochreiter & Schmidhube 提出的。随着 2012 年深度学习的兴起，LSTM 经历了一系列的发展，最终形成了一个相对系统和完整的 LSTM 框架，并在多个领域得到了广泛的应用。LSTM 是一种 RNN 的变种，旨在解决 RNN 在处理长序列数据时遇到的梯度消失和梯度爆炸问题，通过采用遗忘门、输入门和输出门控机制来实现这一目标。LSTM 的网络结构包括一个或多个 LSTM 单元（也称为细胞），它们可以串联在一起以处理序列数据。每个 LSTM 单元都包含了上述的三个门控机制，以及一个细胞状态和一个隐藏状态。这些门控机制允

许 LSTM 在每个时间步骤上选择性地记忆、忘记和输出信息。这些门的功能与滤波器相似，它们可以通过输入门来更新记忆单元的状态，并通过输出门来控制 LSTM 的输出。而遗忘门则负责对记忆单元内的信息进行评估，并决定是丢弃还是保留，可以选择记住重要信息过滤掉噪声信息以此来减轻记忆负担。LSTM 神经网络与循环神经网络在结构上有许多相似之处，但其独特之处在于其内部的各个部分是相互协作的，这有助于更精确地捕获序列中的长期依赖关系。

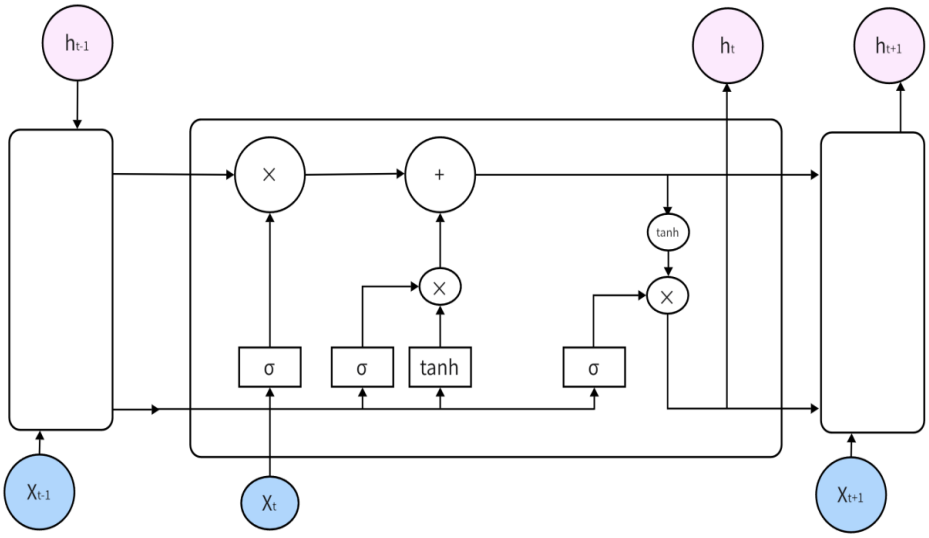


图 3.2 LSTM 单元结构

输入门、遗忘门、输出门的输入数据来源于当前时间段的输入X_t以及前一个时间段的隐藏状态，通过 σ 函数作为激活函数计算得到。假设时间步 t 的小批量输入$X_t \in R^{n\times d}$ 以及前一个时间步输出中的隐藏状态；W_{xx}与b_x为各个门对应的权各门的权重和偏差参数，计算方式如下：

$$I_t = \sigma(X_t W_{xi} + b_i + H_{t-1} W_{hi}) \quad (3\text{-}1)$$

$$F_t = \sigma(X_t W_{xf} + b_f + H_{t-1} W_{hf}) \quad (3\text{-}2)$$

$$O_t = \sigma(X_t W_{xo} + b_o + H_{t-1} W_{ho}) \quad (3\text{-}3)$$

计算之后，需要再计算候选记忆细胞$\tilde{C}_t$，并使用 $tanh$ 函数作为激活函数进行计算：

$$\tilde{C}_t = tanh\,(X_t W_{xc} + b_c + H_{t-1} W_{hc}) \quad (3\text{-}4)$$

LSTM 作为 RNN 的衍生网络，其最大的弱点是运行速度缓慢，主要问题是其前后隐藏状态的依赖性，这使得它无法实现并行处理。

3.2 Transformer 模型

Transformer 由谷歌团队在论文《Attention is All You Need》（2017）提出，是一种用于处理序列数据的架构，特别适用于自然语言处理任务，如翻译、语言生成、问答系统等。

Transformer 模型的一个显著优点是它可以有效地捕获长距离的依赖关系，而无需依赖传统的递归或卷积序列模型。它引入了自注意力机制（self-attention），使得模型可以同时考虑输入序列中的所有位置，从而更好地理解上下文关系。Transformer 模型具有可并行计算的能力，因为它可以在整个序列上进行并行计算，而不需要按顺序处理每个位置。这使得 Transformer 在处理大规模数据时具有较高的效率，改进了 RNN/LSTM“训练慢”的缺点。除了 NLP 任务，Transformer 模型还可以应用于其他序列数据的建模和处理，如音频处理、时间序列预测等。它的灵活性使得它成为处理序列数据的重要工具之一。

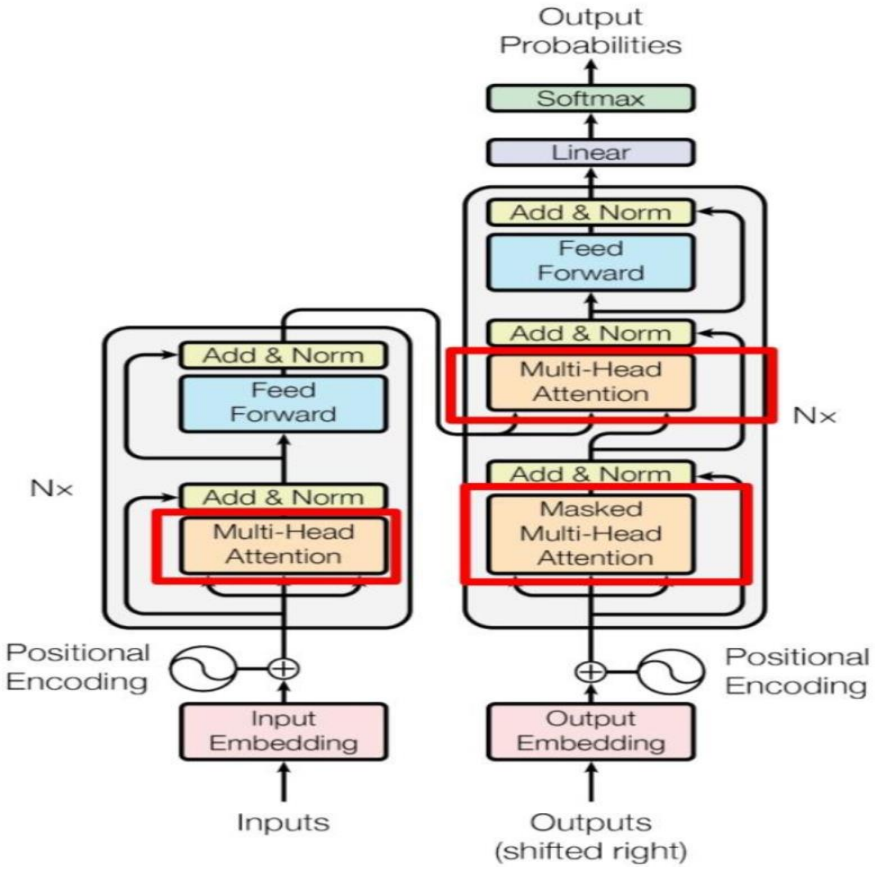


图 3.3 Transformer 模型结构

Transformer 模型采用 Encoder-Decoder 的架构，它是一个基于自注意力机制

（self-attention）的深度学习模型，主要用于处理连续的数据序列。左侧为 Encoder block，右侧为 Decoder block。图中画红色框中的结构为 Multi-Head Attention（多头注意力机制），是由多个 Self-Attention 组合而成的。从 Encoder block 可以观察到一个 Multi-Head Attention，而解码单元框架包含两个多头注意力机制，其中一个与 Masked 有关。在 Multi-Head Attention（多头注意力机制）的上方，还有一个 Add&Norm 层，其中 Add 代表残差连接（Residual Connection），其作用是防止运行的过程中可能造成的网络退化，而 Norm 代表 Layer Normalization，其作用是对每一层的数据值进行标准化。

3.2.1 Transformer 的输入

Transformer 中单词的输入表示 x 由单词 Embedding 和位置 Embedding 相加得到。

单词的 Embedding 在获取方式上可以采用多种途径，比如可以通过 Word2Vec 等算法进行预训练，或者可以直接在 Transformer 中进行训练。位置 Embedding 是用来表示单词在句子里所处的具体位置。通过计算单词与词间关系和相邻两个单词之间距离来确定。Transformer 在实际运行的过程中不利用 RNN 的框架，而是依赖于全局的信息，它不能使用单词的排序信息，这部分信息对于 NLP 是至关重要的。因此，在 Transformer 中，位置 Embedding 的主要作用是被用来存储单词在序列里的相对或绝对位置。当输入一个新单词时，就会出现该单词的位置信息，这个单词就是我们需要保留的单词。位置 Embedding 是用 *PE* 来表示的，既可以通过培训获得，也可以采用特定的数学公式来计算。对于不同类型和规模的单词串来说，其计算方法有所不同。计算公式:

$$PE_{(pos,2i)} = sin\left(po/10000^{2i/d}\right) \tag{3-5}$$

$$PE_{(pos,2i+1)} = cos\left(po/10000^{2i/d}\right) \tag{3-6}$$

其中，*po* 表示单词在句子中的位置，*d* 表示 *PE* 的维度（与词 Embedding 一样），$2i$表示偶数的维度，$2i+1$表示奇数维度（即 $2i \leq d$, $2i+1 \leq d$）。

3.2.2 自注意力机制

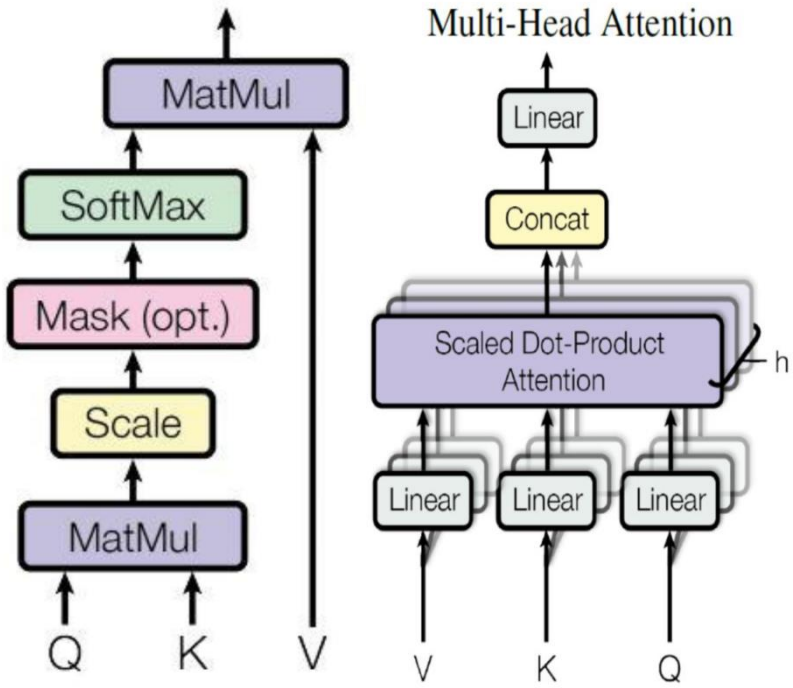


图 3.4 Self-Attention 结构和 Multi-Head Attention

图 3.4 中，左图是 Self-Attention 的结构。受限于现有的计算和存储技术，仅仅通过增加神经元的数量来提高模型的性能是比较困难的，而且过多的参数可能会对模型的整体性能造成不利影响。因此，如何减少神经网络中神经元个数以提高系统的效率就成为一个值得研究的问题。为了解决这一难题，我们可以参考人脑的工作原理。将大脑中存储的知识以某种方式组织起来，然后对这些知识进行处理并形成新的规则，这样就得到了一种新型的神经网络——注意力网络。当人脑集中注意力于关键信息时，也能排除不相关的信息，从而大大提升工作效率，这就是我们常说的注意力机制。换句话说，通过运用注意力机制，该模型可以避免关注所有区域的数据特性，而只需专注于我们感兴趣的区域的数据特性，这样可以减轻计算的负担，提升模型的准确度和泛化能力。Transformer 的核心在于自注意力机制，这使得在处理数据序列时，模型能够同时考虑到输入序列中的每一个位置。传统的序列模型一般采用固定窗口或滑动窗口来捕获上下文关系，但是自注意力机制能够根据输入序列的不同部分自动调整权重，从而更好地捕捉长距离的依赖关系。

若输入序列为$X=\{x_1,x_2,\ldots,x_n\}$，则计算注意力分布公式如下：

$$\alpha_i = softmax(h(x_i,p)) \quad (3\text{-}7)$$

上式中α_i满足α_1+α_2+…+α_n=*1*，表示x_i分布权重：

$$H = att(X,p) = \sum_{t=1}^{n} \alpha_i x_i \quad (3\text{-}8)$$

Transformer 模型也融入了多头注意力机制（如右图所示）。Multi-Head Attention 由多个 Self-Attention 层组成。首先，我们将输入 *X* 传递到 *h* 个不同的 Self-Attention 中，并据此计算出 *h* 个输出矩阵 *H*，允许模型在不同的表示子空间中学习多个不同的注意力表示。这有助于模型更好地捕捉不同类型的关系和特征。

单头自注意力机制首先将输入变量 *X* 进行三轮不同的线性转换，从而产生查询矩阵 *Q*，键矩阵 *K*，值矩阵 *V*。公式如下：

$$Q = Linear(X) = XW^Q \quad (3\text{-}9)$$

$$K = Linear(X) = XW^k \quad (3\text{-}10)$$

$$V = Linear(X) = XW^v \quad (3\text{-}11)$$

接下来，我们使用标准化的点积方法来获得注意力的输出，公式如下：

$$head = attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3\text{-}12)$$

为了增强计算的速度，确保模型能够综合考虑所有相关的信息。在分析了影响模型精度的因素后提出一种新方法，将其转化为一个二次规划问题。因此，我们对输入矩阵 *X* 进行了多次线性映射，得到了 *Q*、*K*、*V* 矩阵，然后将输出的注意力结果进行了拼接，最终实现了提升模型性能的目的。公式如下：

$$MultiHeadAttention = concat(head_1,\ldots,head_h)W^o \quad (3\text{-}13)$$

总体而言，Transformer 模型最显著的优势在于具备处理长距离依赖关系的能力、强大的并行计算性能以及多头注意力机制等特质，这使得它成为处理序列数据的一个关键模型。

3.3 trans_LSTM 融合模型的设计

trans_LSTM 融合模型结构通常是指将 Transformer 和 LSTM（Long Short-Term Memory）这两种神经网络结构进行组合的混合模型。这样的组合可以结合 Transformer 的注意力机制和 LSTM 的长短时记忆，以捕捉序列数据中的不同类型的依赖关系。

Transformer采用了自注意力机制来捕获输入序列中各个位置的相互依赖性，这样在处理序列数据的过程中，就能专注于不同位置的信息，而无需依赖于逐步进行的顺序处理。通常包含多个注意力头，每个头都学习捕捉序列中的不同依赖关系，有助于模型更全面地理解输入序列。在每个注意力层之后，通常会有一个全连接的前馈神经网络。这有助于捕捉非线性关系。但 Transformer 模型并没有天然地处理输入序列中的位置信息，它仅仅通过添加位置编码来引入位置信息，而这种方法对于长序列的处理可能不够精确，因为位置编码的有限表示能力可能无法捕捉复杂的位置关系，比如相对位置信息在自注意力机制处线性变化后会消失。

LSTM 专门设计用于解决梯度消失问题。为了有效地应对序列学习中的长期依赖关系，LSTM 引入了细胞状态和门控机制。这一设计使得 LSTM 能够更有效地捕捉和记忆序列中的短期依赖关系，从而在处理时间序列数据时表现出色。细胞状态允许 LSTM 网络在不同时间步骤之间传递信息，而门控机制则允许网络在不同时间步骤中选择性地记忆或遗忘信息。遗忘门负责确定前一状态中的哪些信息应该被遗忘，输入门负责确定哪些新的信息应该被加入，而输出门则控制最终输出的信息。这种机制使得 LSTM 能够更灵活地处理序列中的各种模式，包括局部模式和短期变化，同时避免了传统 RNN 中容易出现的梯度消失问题。

总体而言，LSTM 在序列建模中的优越性使其成为处理时间序列数据、自然语言处理等领域的重要工具，为深度学习在具有时序特性的任务中的成功应用提供了关键支持在时间序列任务中，通常会面临短时依赖和长时依赖两个方面的挑战。如果未来数据主要受到较为长时刻历史数据的影响，我们将其称为长时依赖。相反，如果未来数据主要受到较为短时刻历史数据的影响，我们则称之为短时依赖。考虑到金融数据的特性，为了解决 Transformer 在捕捉复杂位置关系方面的不足，我们首先使用正弦和余弦函数对输入数据进行编码，然后将编码后的特征输入到自注意力层进行并行特征提取，接着使用 LSTM 进行初步的时序特征提取。最终，我们使用全连接层对 LSTM 和自注意力层提取的特征进行了线性转换，从而得到了预测的价差序列。

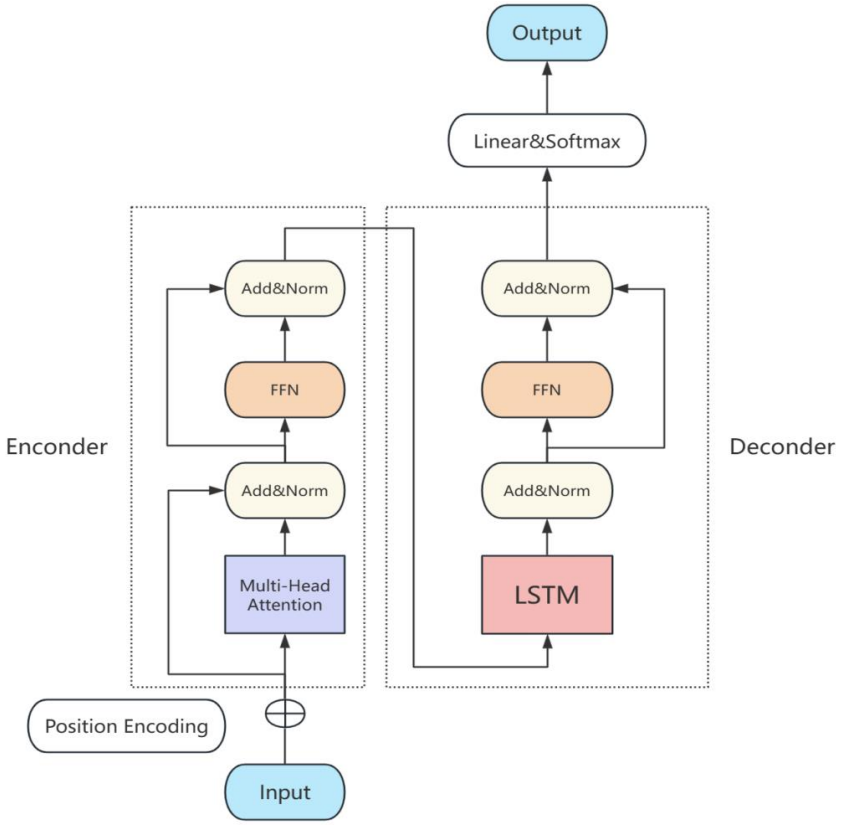


图 3.5 trans_LSTM 融合模型结构

如图 3.5，本文构建 trans_LSTM 融合模型，在输入序列的初步处理中，我们采用了 Transformer 的自注意力机制和多头注意力机制，这两种机制由编码块和解码块构成。每个编码层都包括多头注意力层、全连接层和正则化层，而解码层则由两个多头注意力层组成，以获取全局信息。然后，并在解码器中引入 1 个 LSTM 模块，将这些全局信息传递给 LSTM 层，以便进一步处理并捕捉更长期的依赖关系。这样的结合方式能在处理连续数据时更有效地捕获全局和局部的依赖关系，进而提升模型的整体性能。

具体过程如下：

（1）对于输入的价差数据X_m，分解成片段并在时间维度上关注局部特征，将序列矩阵X_n输入多头注意力机制，得到局部特征表示I_m。

$$I_m = MultiHeadAttention(X_n) \quad (3\text{-}14)$$

（2）分别计算不同时序序列的特征权重S_m，用于计算局部自注意力分数 $H^{(m)}$。

$$S_m = Softmax(ReLU((W_q I_m)^T W_k I_m)) \quad (3\text{-}15)$$

$$H^{(m)} = \sum_{m=1}^{n} S_m W_v I_m \quad (3\text{-}16)$$

（3）对 Transformer 模型输出的特征向量进行了组合，从而形成了特征向量表示 B。

（4）将 B 放入 LSTM 模型中，以便更深入地获取时间序列的相关信息。

$$Y_m^t = LSTM\ (B, W_m, Y_m^{t-1}, \theta_m) \tag{3-17}$$

Y_m^t为 t 时刻的隐藏状态； $LSTM(\bullet)$ 是用于提取 LSTM 模型时序特性的函数；W_m在训练时学习到的一个参数矩阵；θ_m为 LSTM 模型的超参数。

4 基于 trans_LSTM 的套利策略实证分析

4.1 数据收集和清洗

科创板（科技创新板）是中国证券市场的一部分，于 2019 年 7 月 22 日正式成立。科创板的设立旨在支持和促进科技创新企业的发展，为其提供更为便利和灵活的融资渠道。该板块的设立标志着中国资本市场对创新型企业的更加开放和支持，以推动科技创新和经济结构的升级。作为中国新兴股票市场的一部分，科创板具有相对新颖的股票品种和相对较少的市场参与者，这为套利者提供了更为广泛的套利机遇。其独特性质包括新股票、新市场，为投资者带来了更大的不确定性，然而，也正是这种不确定性孕育了更多的投资机会。科创板的股票相对新，市场参与者相对较少，这为套利者提供了更为广阔的操作空间。相对较少的流通股和相对较小的市值，使得这些股票更容易受到市场消息和资金的影响，为套利者提供了更灵活的操作机会。套利者可以通过快速响应市场变化，捕捉到股价波动中的套利机会，实现稳定的收益。此外，科创板股票通常涉及到高科技、创新型企业，这些企业具有较高的成长性和创新性。套利者通过对这些股票的深入研究，可以更好地理解这些企业的技术创新和市场竞争情况，从而把握到套利机会。科技创新的不断推动使得这些股票更容易受到市场情绪和业务动态的影响，为套利者提供了更为敏感的市场操作时机。综合而言，科创板作为新兴市场，其相对新颖的股票和高科技、创新型企业的特性，为套利者提供了更为广泛的套利机会。因此本文选取的数据为科创板数据，数据长度为三个自然年度：2020 年 6 月 1 日-2023 年 6 月 1 日的交易数据，共有 104 支股票，784 个交易日，共计 81536 条数据，并在牛市和熊市两个股票不同状态下回测，数据来源为国泰安数据库。数据在送进模型之前需要进行清洗的步骤，本文中缺失部分主要集中在股票退市和上市不满一年的期间。移除这些部分后，模型的整体学习不会受到影响。因此，在处理缺失值和异常值的数据时，决定直接将其删除。

表 4.1 科创板市场数据描述

	科创板市场原始数据	科创板市场价差数据
最大值	1495.58	1479.11
最小值	2.34	-1489.80
方差	11960.83	16936.97
标准差	109.37	130.14
中位数	49.113	0.89

4.2 数据预处理

本文所使用的深度学习模型是直接对价差数据进行学习，因此首先构建股票见的价差数据：

$$Spread=Close\ price[symbol1]-Close\ price[symbol2] \quad (4\text{-}1)$$

其中，*Close price*表示当日收盘价格，*symbol*表示对应股票。

4.2.1 数据标准化

在进行股票价格预测、风险评估等任务时，使用机器学习模型已成为一种常见的方法。为了提高模型的性能，并促使优化算法更快地达到收敛，标准化是一项关键的预处理步骤，通常在将数据输入模型之前进行。这是因为许多机器学习算法对输入数据的尺度非常敏感，而标准化的过程可以显著减小不同特征之间的权重差异。通过标准化，数据的均值调整为 0，标准差调整为 1，这有助于模型更有效地拟合数据。这样的处理方式有助于防止某些特征在模型训练中占据主导地位，从而使得模型更具鲁棒性。此外，标准化后的数据具有均值为 0、标准差为 1 的特性，这使得数据更容易进行可视化展示。这对于理解数据分布、检测异常值以及直观地观察特征之间的关系都提供了便利。因此，标准化不仅有助于提高模型性能，还为数据的可解释性和可视化提供了有力的支持，使得机器学习在金融领域的应用更为可靠和有效。

$$Spread = \frac{spread-\mu}{\sigma} \quad (4\text{-}2)$$

其中，左侧 *Spread* 为标准化后的数据点，等式右侧 *spread* 为原始数据点，

μ为数据均值，σ为数据方差。

4.2.2 数据转换和模型参数设定

1.数据转换

考虑到 LSTM 模型是监督学习，因此在开始训练前，有必要对时间序列数据进行预处理，将其转化为样本数据和对应的标签数据。LSTM 模型旨在对样本进行非线性拟合，因此问题的性质属于回归问题。为了达到这个目的，采纳了一种适应性强的策略来区分样本数据与标签数据。更具体地说，对序列数据进行了窗口化处理，将从第 t−m 到第 t 分钟的数据标记为样本数据，而将第 t+1 个序列数据视为相应的标签数据。经过使用 m 长度的历史数据进行训练后，最后预测第 t-1 个时间点的数据。这种方法的核心理念是利用过去的数据来预测未来数据点的位置，从而使模型能更有效地捕获序列数据中的非线性关系。整个过程通过对样本数据和标签数据的合理定义，为监督学习问题的建模提供了一种有效而灵活的方法。

2.模型参数设定

本文采用了 trans_LSTM 融合模型和 LSTM 模型。在 LSTM 神经网络模型中，我们将隐含层分为两个层次，并在每一个隐含层的后面都增加了一个 Dropout 层以避免过度拟合，最终还增加了一个全连接层作为输出层。在本研究中，LSTM 神经网络模型的超参数包括批量大小（batch_size=32）、训练轮数（epochs=200）、时间步长（look_back=10）、学习率（learning_rate=0.001），以及两个隐含层上的神经元数和两个 Dropout 层的丢弃率。

4.3 套利策略

初始资金设定为 10000 元，假定 A、B 两个配对资产不需要手续费和印花税，每天进行交易，交易的最小单位是 100 股。首先进行模型初始化，采用随机交易策略，然后每天进行一次训练。要验证两只股票的是否存在长期关系，通常将这两只股票价差的均值作为配对交易的标准，即当股票价差偏离标准时进行相应的后续买卖操作。在实证中，为了判断两只股票价格的高低，通常需要将价差数据进行去中心化处理，从而实现更为高效的分析和预测。去中心化公式：

$$M\_spread_t = spread_t - mean(spread_t) = y_t - \beta x_t - \alpha \quad (4\text{-}3)$$

若两只股票去中心化处理后的价差序列为M_spread_t，该序列满足正态分布，设该正态分布的均值为μ，标准差为σ。具体的交易规则模型如下：

（1）建仓：

当M_spread_t>$\mu+k\sigma$（$k>0$）时，A 被高估，按照 α：1 的比例，买入 B 股票；

当M_spread_t< $\mu-k\sigma$（$k>0$）时，A 被低估，同样按照 α：1 的比例，买入 A 股票。

（2）平仓：

当M_spread_t回到均值 μ 附近时，针对上述建仓操作，反向平仓。

（3）止损：

当M_spread_t>$\mu+\Gamma\sigma$（$\Gamma>k>0$）时，即过分高估 A 股票，执行强制平仓；

当M_spread_t<$\mu-\Gamma\sigma$（$\Gamma>k>0$）时，即过分低估 B 股票时，同样执行强制平仓。

为避免大多数价差落在区间内，导致产生的信号少，同时避免大多数落在区间外过多价差导致交易信号过多，产生高额的交易费用，因此将模型建仓基础信号设置为 $\mu\pm1\sigma$，平仓信号设置为 $\mu\pm0.2\sigma$，止损信号设置为 $\mu\pm2\sigma$。

4.4 评价指标

1.预测指标

平均绝对值误差（MAE）是用来表示预测值与观测值之间绝对误差的平均数。MAE 被定义为一个线性的分数，其中每个个体的差异在其平均值上所占的权重都是一致的。具体公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4\text{-}4)$$

均方误差（MSE）是一个用于衡量回归模型预测值与实际观测值之间偏差的度量标准。MSE 的数值越低越有利，这意味着模型预测的数值与真实数值之间的差距越小。具体公式如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4\text{-}5)$$

均方根误差（RMSE）是用来描述预测值与观测值之间的偏差（也被称为残

差）的样本的标准偏差。均方根误差被用来描述样本的离散性。在进行非线性拟合的过程中，RMSE 的值越低越为理想。具体公式如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4\text{-}6)$$

其中，y_i是实际观测值 ，$\hat{y}_i$是模型预测的值，n 是样本数量。

决定系数（R^2）用于量化回归模型在解释观测数据方差方面的准确性。在大多数情况下，决定系数 R^2 是用来描述模型对数据变异的解释水平，当 R^2 趋近于 1 时，这意味着模型对数据的匹配度更高。公式：

$$SSR = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2 \quad (4\text{-}7)$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4\text{-}8)$$

$$SST = \sum_{i=1}^{n}(\bar{y} - y_i)^2 \quad (4\text{-}9)$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (4\text{-}10)$$

其中，$\bar{y}$表示 y 的平均值，*SSE* 则是模型预测值与实际值差异的平方和，用以表示回归偏差。*SSR* 则代表回归平方和，这是一个反映自变量与因变量相关性的偏差平方和。*SST* 代表的是总平方和，这是观测值与观测均值差异的平方和，用以表示偏差的平方和。

2.策略指标

年化收益率是用于衡量投资或资产组合在一年内表现的指标。它通过将投资的总收益率调整为每年的百分比来表示。年化收益率的计算方法取决于投资的类型和持有期。具体公式如下：

$$Total\ Annualized\ Returns = R_r = ((1+r)^{\frac{250}{n}} - 1) * 100\% \quad (4\text{-}11)$$

其中 r 为策略收益，n 为策略执行天数。这个公式基于假设收益率在持有期内是稳定的。年化收益率允许投资者更容易比较不同投资之间的表现，并提供一个标准化的度量，使得投资者能够更好地评估其投资组合的表现。

最大回撤率定义为一段时期内，投资组合或资产净值从最高点降至最低点的最大百分比。这一指标用于描述可能出现的最糟糕或最极端的亏损状况，并用于评估资产或投资组合在历史数据中可能遭遇的最大损失。评估最大回撤率的能力主要体现在它能够揭示投资组合或资产在特定时间段内的波动和风险程度。较大

的最大回撤率意味着资产或投资组合在历史上曾经经历过较大的损失，可能表明较高的风险水平。具体公式如下：

$$Max\ Drawdown = \frac{Max(P_x - P_y)}{P_x} \quad (4\text{-}12)$$

其中，P_x P_y代表策略某一天的证券与现金的整体价值，y > x。

4.5 回测

4.5.1 科创板套利结果分析

通过对 2020 年 6 月 1 日-2023 年 6 月 1 日的科创板交易数据进行梳理，获得了 784 个交易日、104 支股票的交易数据，使用 Matlab2016 进行价差计算，获得 38329 组股票对的价差数据。将价差数据送入 trans_LSTM 融合模型构架中，为讨论不同股票对的情况，根据按股价关联性程度，最终选取三组股票对进行接下来的分析与讨论。本文选用 LSTM 模型和统计方法中的常用模型-协整模型与 trans_LSTM 融合模型进行比较分析。

在牛市和熊市两种股票情况下分别进行套利策略交易。在金融学领域，“牛市”和“熊市”是两个重要且广泛运用的术语，用以刻画股市和金融市场的整体走势。牛市（Bull Market）标志着市场正处于上升趋势，反映了投资者对未来经济展望的乐观情绪，导致股票和其他资产普遍呈现上涨趋势。这一时期，积极的经济数据和企业盈利增长成为市场主导因素，伴随着投资者信心的显著提升。在牛市中，投资者往往展现出更为敢于承担风险的特质，倾向于采用更积极主动的投资策略，以谋求更高的回报。相较之下，熊市（Bear Market）则反映了市场整体呈下降趋势的状态。在熊市中，投资者对未来经济前景持悲观态度，导致股票和其他资产价格普遍下滑。熊市可能伴随着经济衰退、不利的经济指标、企业盈利的减少以及投资者信心的下滑。在这个阶段，投资者往往更为谨慎，可能选择更为保守的投资策略，并采取避险措施以规避市场风险。牛市和熊市表示这股票市场在不同阶段的不同情况，有助于投资者根据不同的市场环境调整其投资组合以及可以使得投资者更全面地理解市场行为和资产价格波动的原因，这两种情况对投资者制定投资策略和风险管理是至关重要，因此本文选用牛市和熊市环境为

代表进行套利策略交易进一步的讨论。通过对股票市场的分析和观察，最终选择确定牛市回测时间为 2021 年 3 月 1 日-2021 年 6 月 1 日，熊市回测时间为 2022 年 6 月 1 日-2022 年 9 月 1 日。

1.牛市

牛市选择的这三组股票对分别是：聚辰股份（688123）与美迪西（688202）、威胜信息（688100）与晶峰明源（688368）、南微医学（688029）与八亿时空（688181）。

在进行协整检验时，我们观察到上述三组股票对原始股价序列呈现不稳定的趋势。通过将数据进行对数处理后，进行一阶差分，结果显示发现差分序列变得平稳，通过 ADF 检验，即为一阶单整。在进行协整检验时，我们对残差序列进行平稳性检验，均可以通过残差平稳性检验；但在接下来的误差修正模型得到的结果上来看，拟合效果并不好，可以说明序列间存在非线性的关系，使用统计方法的协整模型并不能探究两组序列之间的变化趋势。

表 4.2 协整检验

	688123-688202	688100-688368	688029-688181
ADF 检验	非平稳-非平稳	非平稳-非平稳	非平稳-非平稳
一阶差分序列平稳性	平稳-平稳	平稳-平稳	平稳-平稳
残差平稳性	平稳	平稳	平稳
误差修正模型 mse	6630.9121	682.0351	306.0217
误差修正模型 mae	74.2715	20.5340	14.2047
误差修正模型 r^2	10.8164	3.3522	1.40642

基于以上结果，我们得出结论，若采用基于协整理论考虑统计套利问题时，上述所示的三组股票对的拟合结果价差，在使用协整进行套利策略的研究中都将被排除在外。这将导致投资者在进行投资活动时的选择减少，同时投资风险相对增加。值得注意的是，这一结论的得出是建立在对股价序列的平稳性和协整性的深入研究之上的。

然而，引入 trans_LSTM 融合模型则为我们提供了一种新的视角。通过遍历所有股票对的价差序列，该模型通过对价差数据的可预测性和准确性的判断，发现了基于协整理论未曾发现的套利机会。这表明，传统的协整检验可能未能充分

挖掘市场中隐藏的套利潜力，而新型融合模型的引入为投资者提供了更广泛的选择空间和更准确的套利机会预测，有望在投资决策中发挥重要作用。

其中，价差预测图中，红色曲线为 trans_LSTM 融合模型价差预测、绿色曲线 LSTM 模型价差预测、黑色曲线为实际价差；净值曲线预测图中，红色曲线为 trans_LSTM 融合模型的净值曲线，蓝色曲线为 LSTM 模型的净值曲线。

（1）聚辰股份（688123）与美迪西（688202）股票对

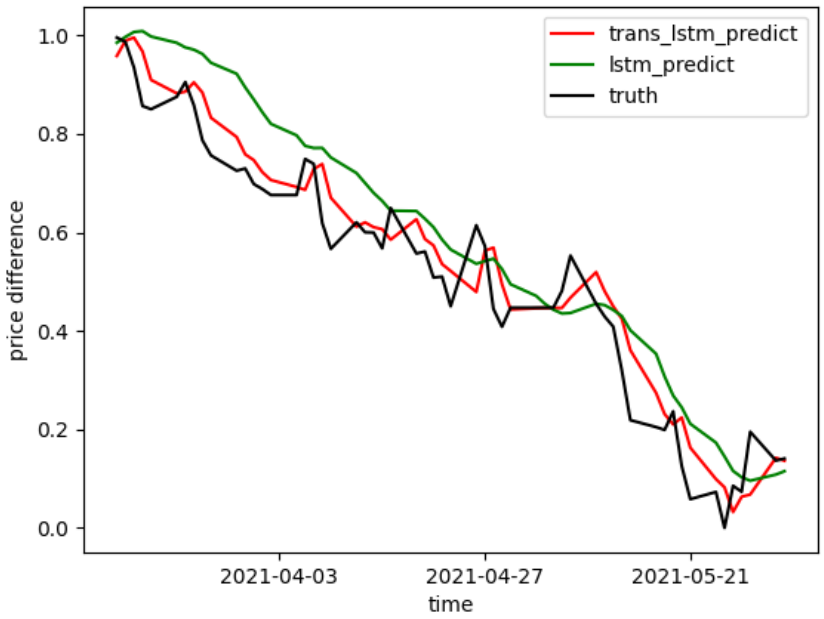


图 4.1 688123-688202 价差预测值和实际值

表 4.3 688123-688202 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0520	0.0042	0.0658	0.9367
LTSM	0.0907	0.0114	0.1069	0.8298

图 4.1 为聚辰股份（688123）与美迪西（688202）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.3 为聚辰股份（688123）与美迪西（688202）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0520 明显小于 LSTM 的 0.0907，从 MSE 来看 trans_LSTM 融合模型为 0.0042 明显小于 LSTM 的 0.0114，从 RMSE 来看 trans_LSTM 融合模型为 0.0658 明显小于 LSTM

的0.1069，说明trans_LSTM融合模型明显优于LSTM模型；从R^2来看trans_LSTM融合模型为0.9367明显大于LSTM的0.8298，说明trans_LSTM融合模型明显优于LSTM模型，综上，对于聚辰股份（688123）与美迪西（688202）股票对的价差预测，trans_LSTM融合模型明显优于LSTM模型。

下面我们将基于这两种模型的预测结果根据套利策略进行交易。

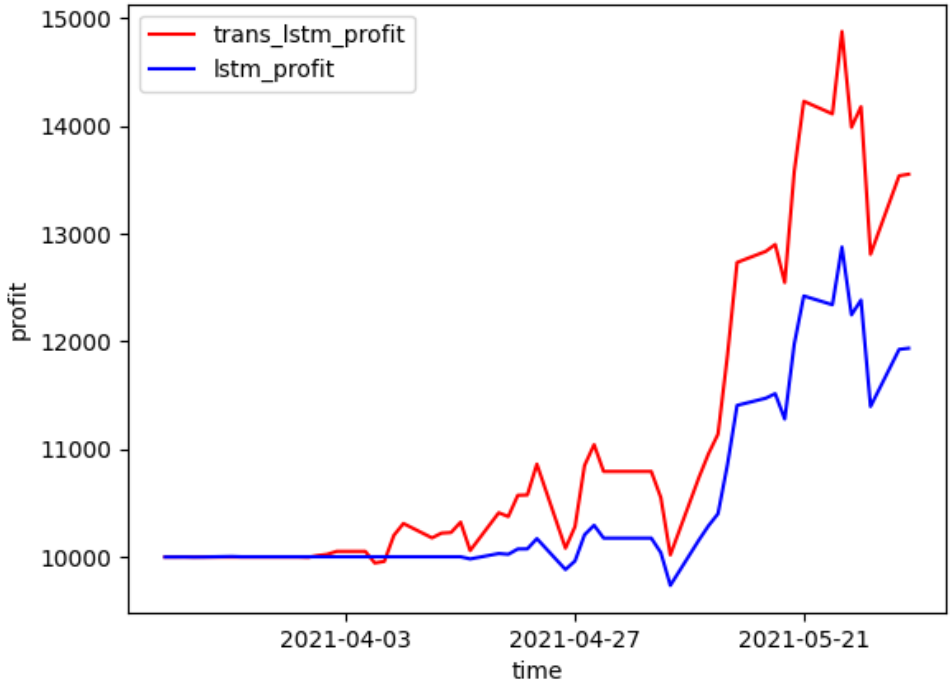


图 4.2 688123-688202 收益净值对比曲线

图4.2为聚辰股份（688123）与美迪西（688202）股票对套利策略收益净值对比曲线，整体呈现上升趋势，并且trans_LSTM融合模型的收益表现得相对较好。在测试阶段前期较为稳定，中间出现了短暂的下降，说明此时套利策略可能由于价差变化存在短暂的波动和偏离；但随后恢复上升的态势，表示在新的市场环境下找到了潜在的机会或者是使用了新的市场条件，两只股票序列的价差逐渐恢复稳定状态，策略净值曲线呈现上升的态势，符合当阶段牛市市场的现状。

接下来将基于trans_LSTM融合模型和基于LSTM模型的预测结果根据设置的套利交易策略进行相同的交易过程并进行对比分析。

表 4.4 688123-688202 回测交易结果

	trans_LSTM	LSTM
回测时间	2021 年 3 月 1 日-2021 年 6 月 1 日	
初始保证金	10000	
交易次数	48	40
总盈利（元）	13551.2090	11936.5170
最大回撤	0.1390	0.1149
年化收益率	2.7057	1.1447

从表 4.4 的交易回测结果来看，在 trans_LSTM 融合模型的基础上，最大的回撤率达到了 0.1390；依据 LSTM 模型计算得出的最大回撤比率为 0.1149。采用 trans_LSTM 融合模型进行的交易次数达到了 48 次；采用 LSTM 模型进行的交易次数达到了 40 次。采用 trans_LSTM 融合模型得出的年化回报率达到了 2.7057；采用 LSTM 模型计算得出的年化回报率达到 1.1447。考虑到最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略收益率明显高于 LSTM 模型，并且与当前的牛市趋势相吻合。

（2）威胜信息（688100）与晶峰明源（688368）股票对

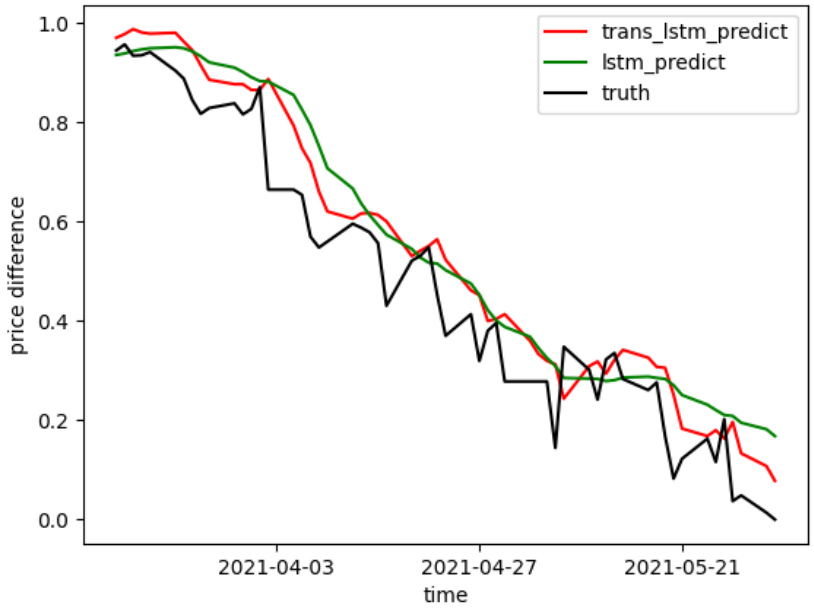


图 4.3 688100-688368 价差预测值和实际值

表 4.5 688100-688368 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0712	0.0077	0.0880	0.9041
LTSM	0.0823	0.0107	0.1036	0.8670

图 4.3 为威胜信息（688100）与晶峰明源（688368）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.5 为威胜信息（688100）与晶峰明源（688368）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0712 明显小于 LSTM 的 0.0823，从 MSE 来看 trans_LSTM 融合模型为 0.0077 明显小于 LSTM 的 0.0107，从 RMSE 来看 trans_LSTM 融合模型为 0.0880 明显小于 LSTM 的 0.1036，从 R^2 来看 trans_LSTM 融合模型为 0.9041 明显大于 LSTM 的 0.8670，说明 trans_LSTM 融合模型明显优于 LSTM 模型，综上，对于威胜信息（688100）与晶峰明源（688368）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

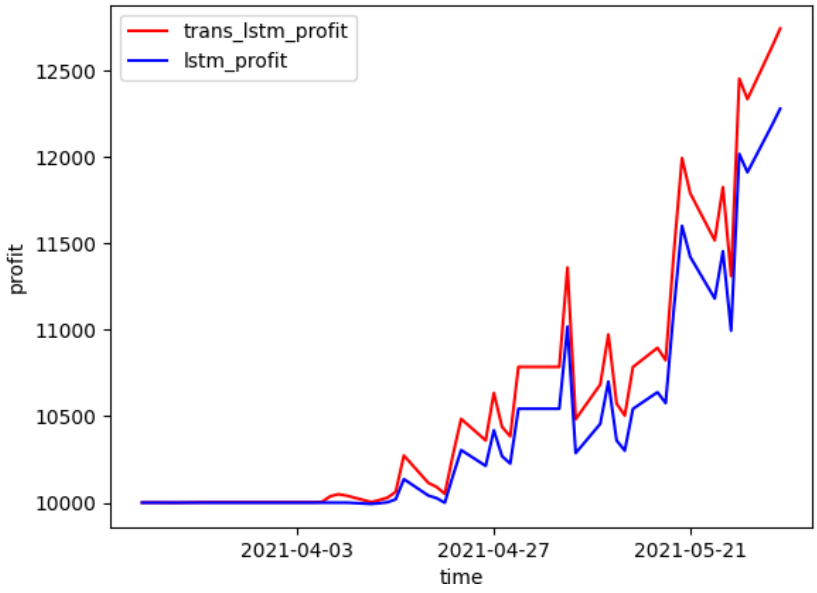


图 4.4 688100-688368 收益净值曲线对比

图 4.4 为威胜信息（688100）与晶峰明源（688368）股票对收益净值对比曲线，整体呈现上升的趋势，并且 trans_LSTM 融合模型的收益表现得相对较好。接下来将基于 trans_LSTM 融合模型和基于 LSTM 模型的预测结果根据设

置的统计套利交易策略进行相同的交易过程并进行对比分析。

表 4.6 688100-688368 回测交易结果

	trans_LSTM	LSTM
回测时间	2021 年 3 月 1 日-2021 年 6 月 1 日	
初始保证金	10000	
交易次数	48	38
总盈利（元）	12743.0660	12279.2250
最大回撤	0.0773	0.0664
年化收益率	1.8429	1.4230

从表 4.6 的交易回测结果来看，在 trans_LSTM 融合模型的基础上，最大的回撤率达到了 0.0773；采用 LSTM 模型的三组股票的最大回撤率为 0.0664。采用 trans_LSTM 融合模型进行的交易次数达到了 48 次；采用 LSTM 模型进行的交易次数达到了 38 次。采用 trans_LSTM 融合模型得出的年化收益率达到了 1.8429；采用 LSTM 模型计算出的年化回报率达到了 1.4230。考虑到最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

（3）南微医学（688029）和八亿时空（688181）股票对

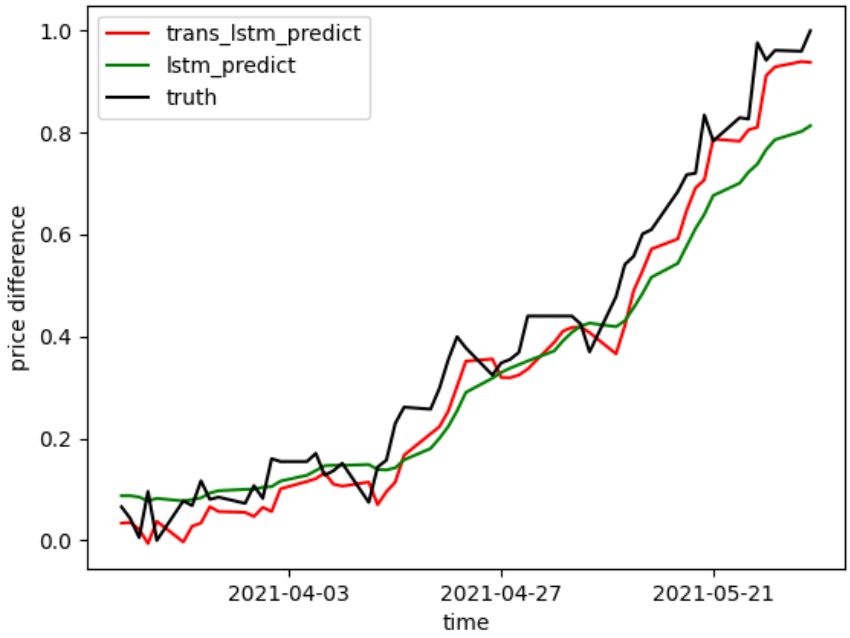


图 4.5 688029-688181 价差预测值和实际值

表 4.7 688029-688181 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0543	0.0042	0.0651	0.9508
LTSM	0.0718	0.0085	0.0923	0.9011

图 4.5 为南微医学（688029）和八亿时空（688181）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.7 为南微医学（688029）和八亿时空（688181）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0543 明显小于 LSTM 的 0.0718，从 MSE 来看 trans_LSTM 融合模型为 0.0042 明显小于 LSTM 的 0.0085，从 RMSE 来看 trans_LSTM 融合模型为 0.0651 小于 LSTM 的 0.0923，从 R^2 来看 trans_LSTM 融合模型为 0.9508 明显大于 LSTM 的 0.9011，说明 trans_LSTM 融合模型明显优于 LSTM 模型，综上，对于南微医学（688029）和八亿时空（688181）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

下面将基于这两种模型的预测结果根据套利策略进行交易。

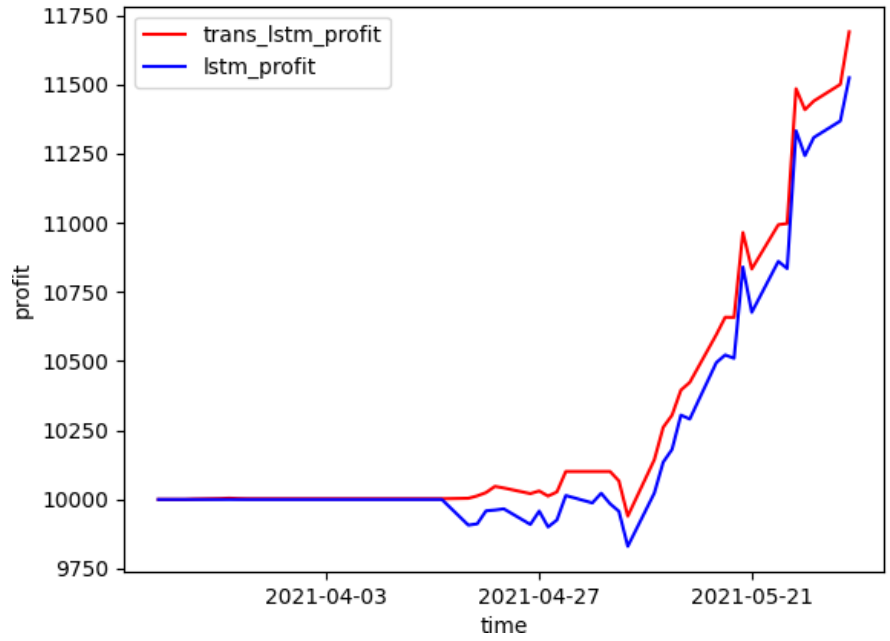


图 4.6 688029-688181 收益净值曲线

图 4.6 为南微医学（688029）和八亿时空（688181）股票对收益净值对比曲线，trans_LSTM 融合模型的收益表现得相对较好。整体来讲是向上的趋势，符

合当前牛市现状。

表 4.8 688029-688181 交易回测结果

	trans_LSTM	LSTM
回测时间	2021 年 3 月 1 日-2021 年 6 月 1 日	
初始保证金	10000	
交易次数	35	32
总盈利（元）	11689.9170	11523.8068
最大回撤	0.0160	0.0191
年化收益率	0.9602	0.8428

从表 4.8 的交易回测结果来看，采用 trans_LSTM 融合模型时，最大的回撤率达到了 0.0160；采用 LSTM 模型的三组股票的最大回撤率为 0.0191。采用 trans_LSTM 融合模型进行的交易次数达到了 35 次；采用 LSTM 模型进行的交易次数达到了 32 次。采用 trans_LSTM 融合模型得出的年化回报率达到了 0.9602；采用 LSTM 模型计算得出的年化回报率达到了 0.8428。考虑到最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

通过价差预测值和实际值对比图可以看到，在三组股票对上 trans_LSTM 融合模型的曲线走势更为贴近实际价差的变化趋势，与真实值更为贴近。通过计算对比分析 MAE、MSE、RMSE 以及 R^2 的数值大小以及交易回测的结果上所得到的策略评价指标年化收益率与最大回撤率的比较上也可以看到本文提出的 trans_LSTM 融合模型的效果和性能与 LSTM 模型相比得到了提升，并且符合牛市和熊市的现状。

2.熊市

同样，挑选出卓越新能（688196）和奥特维（688516）股票对，天准科技（688003）和石头科技（688169）股票对和江苏北人（688218）和建龙微纳（688357）股票对在熊市时间下进行回测对比评价。

表 9 协整检验

	688196-688516	688003-688169	688218-688357
ADF 检验	平稳-非平稳	非平稳-非平稳	非平稳-非平稳
一阶差分序列平稳性	平稳-平稳	平稳-平稳	平稳-平稳
残差平稳性	平稳	平稳	平稳
误差修正模型 mse	566.6223	7234.7546	178.0489
误差修正模型 mae	20.1518	71.7000	12.0261
误差修正模型 r^2	2.1551	6.0207	0.69610

同样，我们得出结论，若采用基于协整理论考虑统计套利问题时，上述所示的三组股票对都将被排除在外。这将导致投资者在进行投资活动时的选择减少，同时投资风险相对增加。

（1）卓越新能（688196）和奥特维（688516）股票对

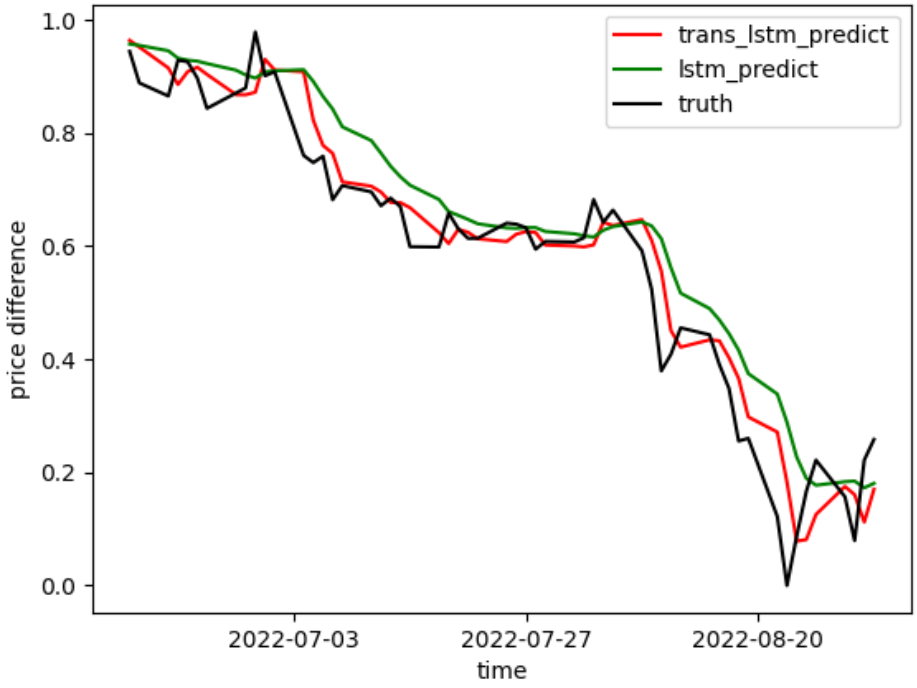


图 4.7 688196-688516 价差预测值和实际值

表 4.10 688196-688516 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0474	0.0043	0.0656	0.9335
LTSM	0.0704	0.0088	0.0941	0.8631

图 4.7 为卓越新能（688196）和奥特维（688516）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.10 为卓越新能（688196）和奥特维（688516）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0474 明显小于 LSTM 的 0.0704，从 MSE 来看 trans_LSTM 融合模型为 0.0043 明显小于 LSTM 的 0.0088，从 RMSE 来看 trans_LSTM 融合模型为 0.0656 明显小于 LSTM 的 0.0941，从 R^2 来看 trans_LSTM 融合模型为 0.9335 明显大于 LSTM 的 0.8631，说明 trans_LSTM 融合模型明显优于 LSTM 模型。综上，对于卓越新能（688196）和奥特维（688516）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

下面我们将基于这两种模型的预测结果根据套利策略进行交易。

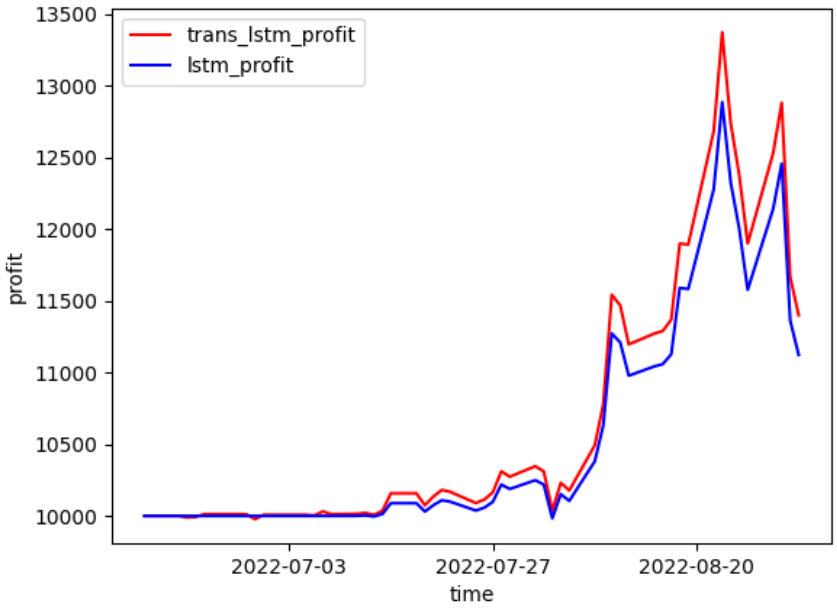


图 4.8 688196-688516 收益净值对比曲线

图 4.8 为卓越新能（688196）和奥特维（688516）股票对套利策略收益净值对比曲线，整体的形式处于动荡波动的状态，可以看到最终的收益水平不是很理

想。

表 4.11 688196-688516 回测交易结果

	trans_LSTM	LSTM
回测时间	2022 年 6 月 1 日-2022 年 9 月 1 日	
初始保证金	10000	
交易次数	47	38
总盈利（元）	11399.6630	11122.5840
最大回撤	0.1475	0.1368
年化收益率	0.7763	0.5946

从表 4.11 的交易回测结果来看，在 trans_LSTM 融合模型的基础上，最大的回撤率达到了 0.1475；采用 LSTM 模型的三组股票的最大回撤率为 0.1368。采用 trans_LSTM 融合模型进行的交易次数达到了 47 次；采用 LSTM 模型进行的交易次数达到了 38 次。采用 trans_LSTM 融合模型，年化收益率达到了 0.7763；采用 LSTM 模型计算出的年化回报率达到了 0.5946。考虑最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

（2）天准科技（688003）和石头科技（688169）股票对

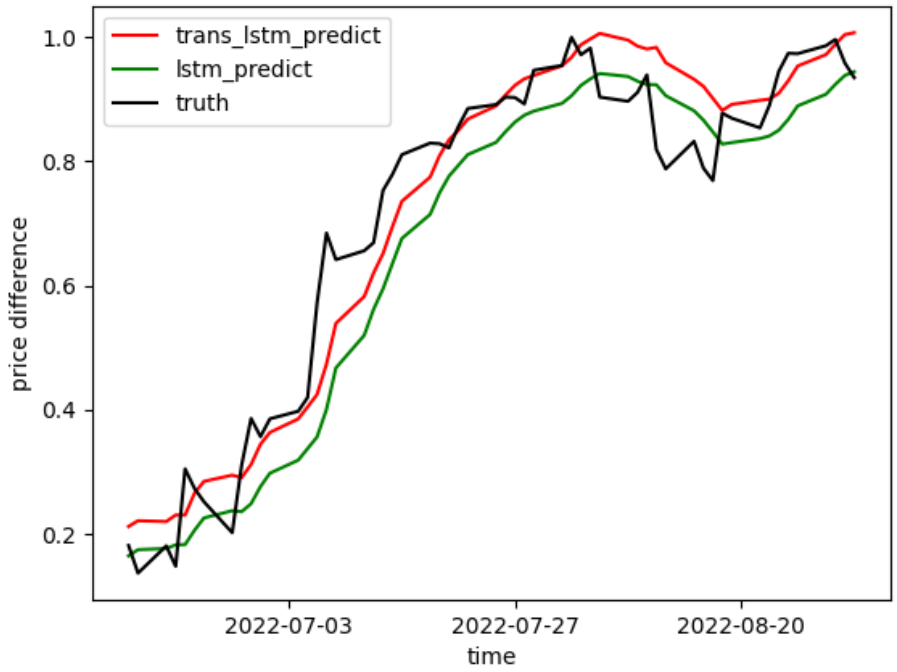


图 4.9 688003-688169 价差预测值和实际值

表 4.12 688003-688169 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0541	0.0053	0.0727	0.9278
LTSM	0.0763	0.0085	0.0924	0.8835

图 4.9 为天准科技（688003）和石头科技（688169）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.12 为天准科技（688003）和石头科技（688169）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0541 明显小于 LSTM 的 0.0763，从 MSE 来看 trans_LSTM 融合模型为 0.0053 小于 LSTM 的 0.0085，从 RMSE 来看 trans_LSTM 融合模型为 0.0727 明显小于 LSTM 的 0.0924，从 R^2 来看 trans_LSTM 融合模型为 0.9278 明显大于 LSTM 的 0.8835，说明 trans_LSTM 融合模型明显优于 LSTM 模型。综上，对于天准科技（688003）和石头科技（688169）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

下面将基于这两种模型的预测结果根据套利策略进行交易。

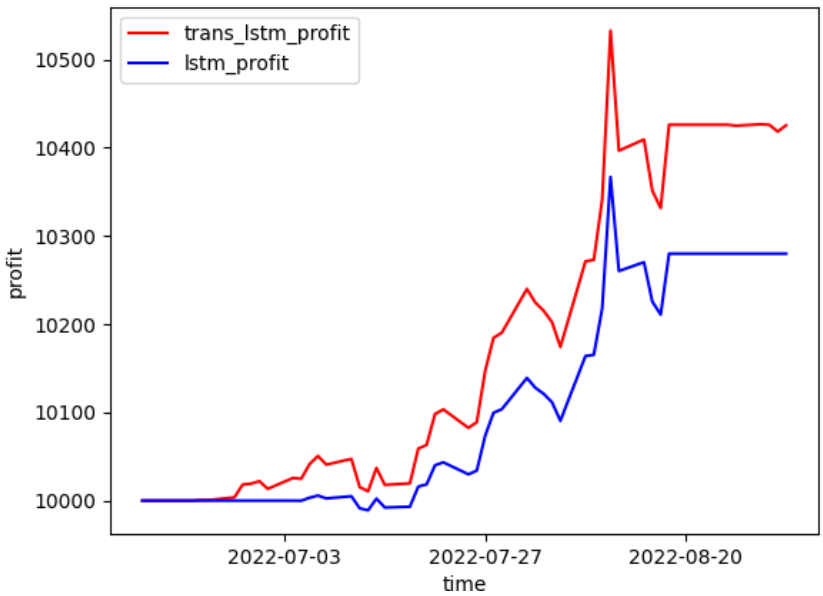


图 4.10 688003-688169 收益净值曲线对比

表 4.13 688003-688169 回测交易结果

	trans_LSTM	LSTM
回测时间	2022 年 6 月 1 日-2022 年 9 月 1 日	
初始保证金	10000	
交易次数	47	30
总盈利（元）	10425.4570	10279.9370
最大回撤	0.0191	0.0150
年化收益率	0.2005	0.1287

从表 4.13 的交易回测结果来看，采用 trans_LSTM 融合模型时，最大的回撤率达到了 0.0191；三组基于 LSTM 模型的股票的最大回撤率为 0.0150。采用 trans_LSTM 融合模型进行的交易次数达到了 47 次；交易次数基于 LSTM 模型设定为 30 次。采用 trans_LSTM 融合模型，年化收益率达到了 0.2005；采用 LSTM 模型计算得出的年化回报率达到了 0.1287。考虑最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

（3）江苏北人（688218）和建龙微纳（688357）股票对

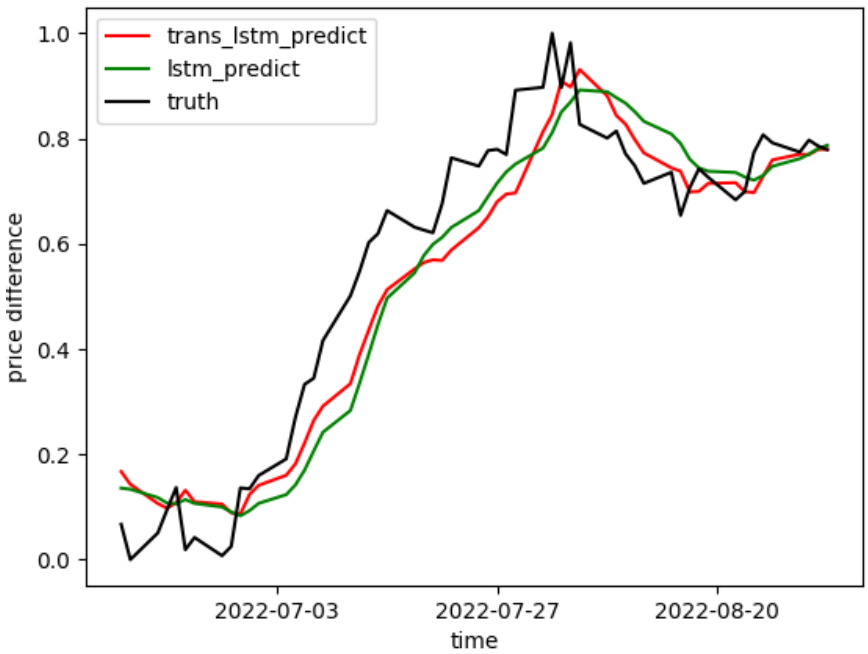


图 4.11 688218-688357 价差预测值和实际值

表 4.14 688218-688357 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0747	0.0082	0.0908	0.9078
LTSM	0.0853	0.0104	0.1019	0.8839

图 4.11 为江苏北人（688218）和建龙微纳（688357）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.14 为江苏北人（688218）和建龙微纳（688357）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0747 明显小于 LSTM 的 0.0853，从 MSE 来看 trans_LSTM 融合模型为 0.0082 明显小于 LSTM 的 0.0104，从 RMSE 来看 trans_LSTM 融合模型为 0.0908 明显小于 LSTM 的 0.1019，从 R^2 来看 trans_LSTM 融合模型为 0.9078 明显大于 LSTM 的 0.8839，说明 trans_LSTM 融合模型明显优于 LSTM 模型。综上，对于江苏北人（688218）和建龙微纳（688357）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

下面将基于这两种模型的预测结果根据套利策略进行交易。

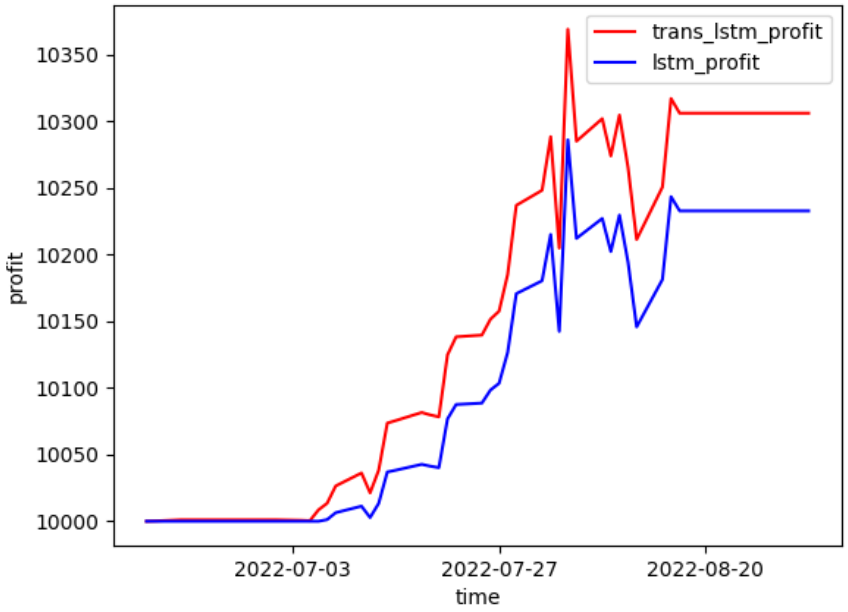


图 4.12 688218-688357 收益净值曲线

图 4.12 为江苏北人（688218）和建龙微纳（688357）股票对收益净值对比曲线，trans_LSTM 融合模型的收益表现得相对较好。接下来将基于 trans_LSTM 融

合模型和基于 LSTM 模型的预测结果根据设置的统计套利交易策略进行相同的交易过程并进行对比分析。

表 4.15 688218-688357 交易回测结果

	trans_LSTM	LSTM
回测时间	2022 年 6 月 1 日-2022 年 9 月 1 日	
初始保证金	10000	
交易次数	32	30
总盈利（元）	10306.1050	10232.8150
最大回撤	0.0152	0.0136
年化收益率	0.1414	0.1064

从表 4.15 的交易回测结果来看，采用 trans_LSTM 融合模型时，最大的回撤率达到了 0.0152；采用 LSTM 模型的三组股票的最大回撤率为 0.0136。采用 trans_LSTM 融合模型进行的交易次数达到了 32 次；交易次数基于 LSTM 模型定为 30 次。采用 trans_LSTM 融合模型得出的年化回报率达到了 0.1414；采用 LSTM 模型计算出的年化回报率达到了 0.1064。考虑最大的回撤和年化收益率，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

通过对比价差预测值与实际值的图表，我们可以观察到，在三组股票的对比中，trans_LSTM 融合模型的曲线趋势更接近实际的价差变化，与实际值更为接近。通过对 MAE、MSE、RMSE 和 R^2 的数值大小以及交易回测结果进行计算和对比分析，我们可以观察到，与 LSTM 模型相比，本研究提出的 trans_LSTM 融合模型在策略评价指标的年化收益率和最大回撤率方面都有了明显的提升。

当我们在牛市和熊市期间进行回测分析时，可以观察到，所获得的总盈利数据与当前市场的实际情况是一致的。在牛市阶段，股票呈现出持续上升的态势，与之相对应的收益回报也相对较高；在熊市阶段，股票市场表现不佳，导致收益相对较低。

4.5.2 跨市科创板和主板套利结果分析

为进一步探究模型的适用性，接下来进行跨市套利策略的分析。跨市套利是

指通过在不同的市场进行交易，利用价格差异或其他市场特征，以获取利润的一种投资策略。这种策略通常涉及在两个或多个市场中同时买入和卖出相关资产，以利用市场之间的差价或价差。最常见的跨市套利类型之一是价差套利，这涉及在两个不同市场上同时操作，以利用价格差异。这可能包括不同市场、不同交易所或不同金融工具之间的价差，投资者可能在不同市场中进行跨市套期保值，以对冲价格波动的风险。投资者通过在高估值市场卖出，低估值市场买入，在这两个不同类型的公司中进行资金分配，实现风险的分散，实现套利机会。主板市场上的公司通常经过较长时间的发展，有一定的规模和市场份额；科创板市场更侧重于创新型企业，而主板市场上的公司相对成熟，并且主板市场上的公司涵盖了各个行业，包括制造业、服务业、金融业等。这使得投资者可以在不同行业中进行选择，实现更好的资产配置。通过对 2020 年 6 月 1 日-2023 年 6 月 1 日的主板交易数据进行梳理，获得了 784 个交易日、1752 支股票的交易数据。根据市场研究和行业分析，选择具有潜力的行业和公司。关注主板市场和科创板市场中相似行业的公司，选择出 200 支主板股票，并结合 104 支科创板股票，最终在牛熊环境下各选一组股票对进行接下来的分析与讨论。牛市股票对为：美迪西（688202）和昊华科技（600378）股票对，熊市股票对为：开普云（688228）和众源新材（603527）股票对。

表 4.16 主板市场数据描述

	主板市场原始数据	科创板-主板价差数据
最大值	19779.82	1492.27
最小值	3.03	-19774.13
方差	1257948.36	1269759.56
标准差	1121.58	1126.84
中位数	31.80	12.63

在进行价差预测和模拟交易之前，同样对选择的股票对进行协整检验，股票原始序列非平稳序列，一阶差分后序列平稳，通过 ADF 检验。进一步在进行协整检验时发现，残差通过平稳性检验的同时误差修正模型的拟合效果并不好，说明序列间存在这非线性的关系。

1.牛市股票对：美迪西（688202）和昊华科技（600378）股票对

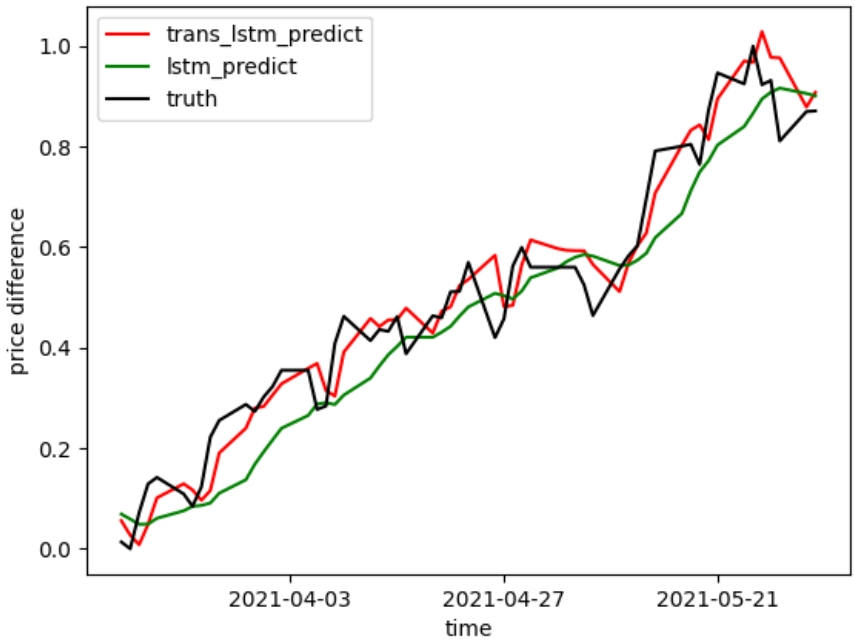


图 4.13　688202-600378 价差预测值和实际值

表 4.17　688202-600378 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0476	0.0036	0.0599	0.9467
LTSM	0.0695	0.0070	0.0831	0.0831

图 4.13 为美迪西（688202）和昊华科技（600378）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.17 为美迪西（688202）和昊华科技（600378）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0476 明显小于 LSTM 的 0.0695，从 MSE 来看 trans_LSTM 融合模型为 0.0036 小于 LSTM 的 0.0070，从 RMSE 来看 trans_LSTM 融合模型为 0.0599 明显小于 LSTM 的 0.0831，从 R^2 来看 trans_LSTM 融合模型为 0.9467 明显大于 LSTM 的 0.0831，说明 trans_LSTM 融合模型明显优于 LSTM 模型，综上，对于美迪西（688202）和昊华科技（600378）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM 模型。

下面将基于这两种模型的预测结果根据套利策略进行交易。

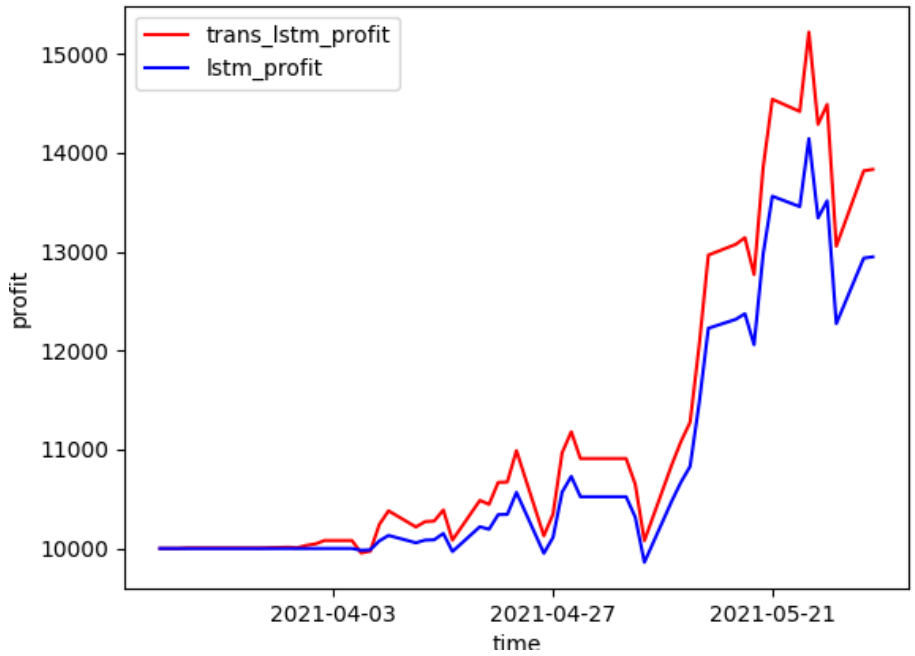


图 4.14　688202-600378 收益净值曲线

图 4.14 为美迪西（688202）和昊华科技（600378）股票对收益净值对比曲线，trans_LSTM 融合模型的收益表现得相对较好。

接下来将基于 trans_LSTM 融合模型和基于 LSTM 模型的预测结果根据设置的套利交易策略进行相同的交易过程并进行对比分析。

表 4.18　688202-600378 交易回测结果

	trans_LSTM	LSTM
回测时间	2021 年 3 月 1 日-2021 年 6 月 1 日	
初始保证金	10000	
交易次数	49	42
总盈利（元）	13833.0655	12948.7500
最大回撤	0.1423	0.1323
年化收益率	3.0496	2.0461

从表 4.18 的交易回测结果来看，在 trans_LSTM 融合模型的基础上，最大的回撤率达到了 0.1423；基于 LSTM 模型，三组股票的最大回撤率为 0.1323。基于 trans_LSTM 的融合模型，交易的次数达到了 49 次；采用 LSTM 模型进行的交易次数达到了 42 次。采用 trans_LSTM 融合模型得出的年化回报率达到了 3.0496；

采用 LSTM 模型计算得出的年化回报率达到了 2.0461。综合对比最大的回撤和年化收益率指标，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

2.熊市股票对：开普云（688228）和众源新材（603527）股票对

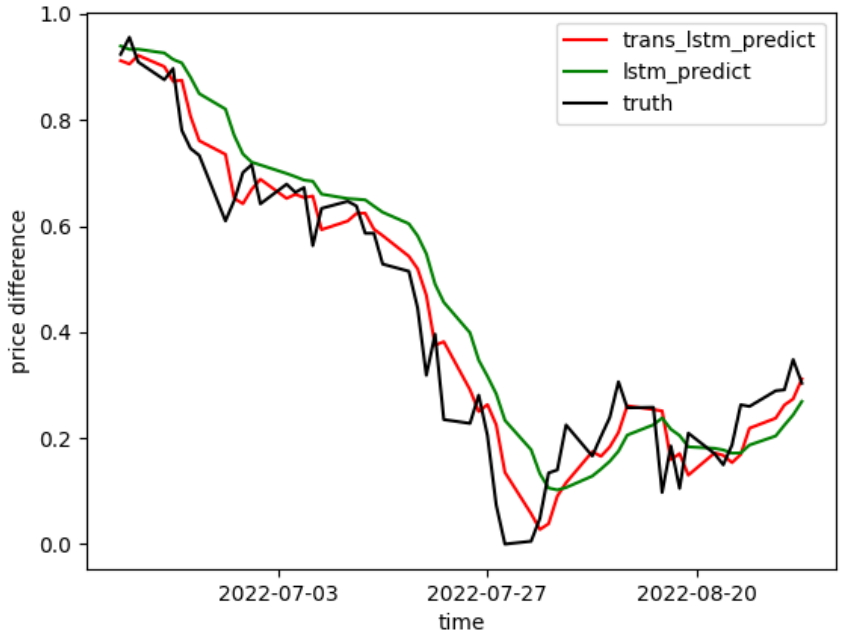


图 4.15　688228-603527 价差预测值和实际值

表 4.19　688228-603527 预测值与实际值的对比

	MAE	MSE	RMSE	R^2
trans_LSTM	0.0524	0.0045	0.0669	0.9368
LTSM	0.0807	0.0103	0.1012	0.8544

图 4.15 为开普云（688228）和众源新材（603527）股票对价差预测值和实际值对比曲线，从曲线上来看 trans_LSTM 融合模型预测结果比 LSTM 模型预测结果更贴近真实值。表 4.19 为开普云（688228）和众源新材（603527）股票对预测值与实际值的结果指标对比，从 MAE 来看 trans_LSTM 融合模型为 0.0524 明显小于 LSTM 的 0.0807，从 MSE 来看 trans_LSTM 融合模型为 0.0045 明显小于 LSTM 的 0.0103，从 RMSE 来看 trans_LSTM 融合模型为 0.0669 明显小于 LSTM 的 0.1012，从 R^2 来看 trans_LSTM 融合模型为 0.9368 明显大于 LSTM 的 0.8544，说明 trans_LSTM 融合模型明显优于 LSTM 模型，综上，对于开普云（688228）和众源新材（603527）股票对的价差预测，trans_LSTM 融合模型明显优于 LSTM

模型。

下面将基于这两种模型的预测结果根据套利策略进行交易。

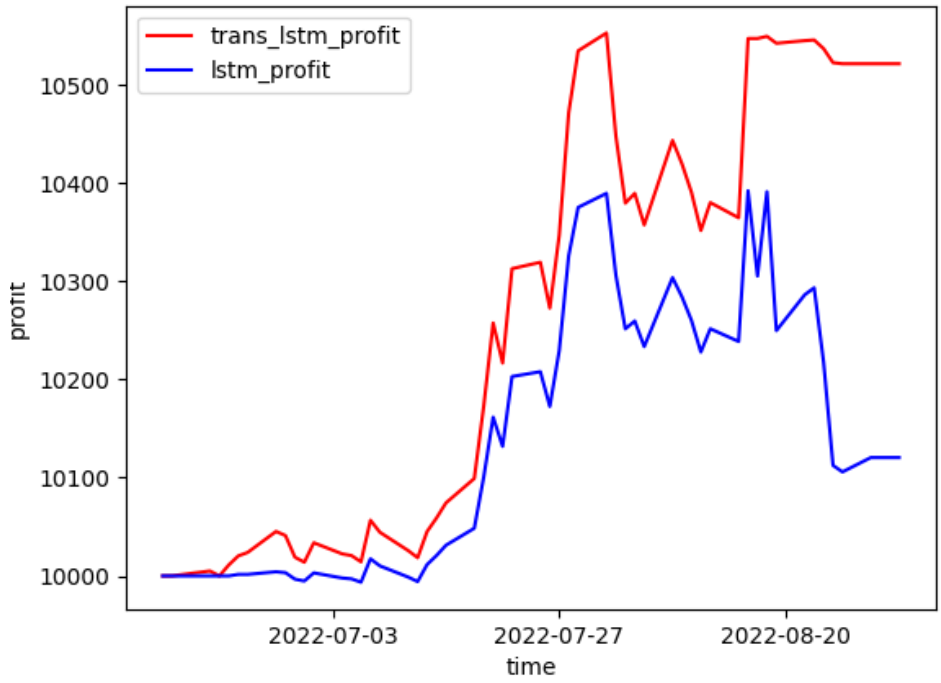


图 4.16　688228-603527 收益净值曲线

图 4.16 为开普云（688228）和众源新材（603527）股票对收益净值对比曲线，trans_LSTM 融合模型的收益表现得相对较好。接下来将基于 trans_LSTM 融合模型和基于 LSTM 模型的预测结果根据设置的统计套利交易策略进行相同的交易过程并进行对比分析。

表 4.19　688228-603527 交易回测结果

	trans_LSTM	LSTM
回测时间	2022 年 6 月 1 日-2022 年 9 月 1 日	
初始保证金	10000	
交易次数	51	47
总盈利（元）	10521.8760	10120.5550
最大回撤	0.0190	0.0275
年化收益率	0.2452	0.0530

从表 4.19 的交易回测结果来看，在 trans_LSTM 融合模型的基础上，最大的回撤率达到了 0.0190；三组基于 LSTM 模型的股票的最大回撤率为 0.0275。采

用 trans_LSTM 融合模型时，交易的次数达到了 51 次；采用 LSTM 模型进行的交易次数达到了 47 次。采用 trans_LSTM 融合模型，年化收益率达到了 0.2452；采用 LSTM 模型计算出的年化回报率达到了 0.0530。综合对比最大的回撤和年化收益率指标，trans_LSTM 融合模型的交易策略所带来的收益率明显高于 LSTM 模型。

同样，模型在跨市套利中分别在牛市和熊市的时间段进行回测分析时可以看出，我们得到了在科创板套利策略情况下相同的结论，即融合模型的评价指标和策略收益更优，trans_LSTM 融合模型的应用得到了扩展，并且根据套利策略所得到的总盈利情况也是符合市场现状的。在牛市时期，应收益回报同样较高；在熊市时期，得到收益较低的结果。

4.6 本章小结

本章首先对于数据的选取和清洗进行说明，并且对数据的转换、模型的参数设定、本文所使用的套利策略以及所使用的评价指标进行了介绍和说明。首先考虑科创板市场内部的套利问题，由于股票之间存在的关联性质以及股票市场自身的特性，最终在牛市和熊市两个市场状态下分别选择出三组股票对进行分析。经过实证分析结果可以看到，这六组股票对的协整拟合的效果并不理想，可以说明基于协整理论不能深入的挖掘股票对之间的深入联系，这可能表明当前发现了基于协整理论未能发现的套利关系；随后通过对比 LSTM 和 trans_LSTM 的各项评价指标结果可以看到本文所提出的 trans_LSTM 融合模型拟合效果和收益方面更好。为了进一步的考虑模型的适用性和避免结果的偶然性，在科创板市场与主板市场之间进行了跨市套利的研究，同样得到了与前文相一致的结论。多次实证结果证明，基于 trans_LSTM 融合模型在分析股票套利问题上相较于基于协整理论和 LSTM 模型上更为稳定且收益更高。

5 结论与展望

5.1 研究结论

本文从传统统计分析方法和机器学习方法两个角度对现有国内外文献进行了梳理，明确了文章的研究方向、研究目的以及研究意义。阐述了套利的基本储备知识，包括套利的定义与类型和统计套利的发展以及主流的两种套利策略的特点，介绍了 LSTM 模型的模型结构、Transformer 模型，以及基于 LSTM 模型和 Transformer 模型各自优势进一步构建了 trans_LSTM 融合模型，最后通过对价差数据的可预测性和策略评价来判断是否存在可套利空间，为深度学习技术在金融领域的进一步发展提供了新方向，提出了新方法。

在构建基于 trans_LSTM 融合模型的套利策略时，首先提出了一种结合 LSTM 和 Transformer 的融合模型来预测价差数据的走势。考虑到 RNN 和其衍生模型在前后隐藏状态上的依赖性，这导致数据在训练的过程中难以并行处理，而使用具备自注意力机制的 Transformer 模型恰好可以稳定地提取序列的长期特征，这使得 LSTM 和 Transformer 模型在提取股票数据特征时展现出不同的趋势，二者的结合增强了对股票价差的预测准确性。然后利用 trans_LSTM 融合模型对科创板数据价差和跨市价差进行预测, 在牛市和熊市环境下，在不同的股票对之间与 LSTM 模型的预测结果进行对比，显示 trans_LSTM 融合模型训练的预测的价差与实际的价差在分布上呈现出更高的相似性。另外与基于统计方法研究套利问题进行对比，以多数学者使用的协整理论为代表，发现了基于协整理论未曾发现的套利机会。通过对比实验，将 trans_LSTM 融合模型与 LSTM 模型进行比较，在不同外部环境下本文提出的融合模型均取得最优结果，证明了 trans_LSTM 融合模型的效果和性能更优，进一步通过跨市套利的结果分析也表明了融合模型在统计套利任务中具有良好的普适性。因此，将深度学习模型应用到统计套利方面有一定的研究价值和应用价值。

总的来说，本文引入了一种基于 trans_LSTM 融合模型的新方法。通过详细的实证验证，本文的深度学习模型在价差预测方面展现出卓越的预测精度，并且具备较强的普适性。这一研究为解决统计套利问题提供了有益的参考，同时也为深度学习方法在金融领域的进一步发展提供了新的思路和可能性。本文提出的

trans_LSTM 融合模型不仅在价差预测上取得了显著的成果，而且在不同环境中均表现出良好的通用性，突显了其对多样化金融数据和复杂市场情境的适应性。这一模型为金融市场中价差变动的准确预测提供了一种有效的工具，并为从中寻找潜在统计套利机会奠定了基础。综上所述，本文所提出的基于 trans_LSTM 融合模型的深度学习方法在金融领域的价差预测问题上取得了较好的效果，为金融领域应用深度学习技术的研究和实践提供了新的探索方向，为解决市场中的统计套利问题提供了有益的启示。

5.2 不足与展望

本研究利用深度学习技术，并使用价差数据来探讨统计套利的问题。但由于研究时间和研究能力的局限性，股票市场的变化趋势研究仍存在一些待完善的地方。为此，我们在下文中详细列举了这些问题，以便后续的研究能够进行有针对性的优化。从数据的角度看，本文选择科创板的交易日度数据作为研究样本，这并不有利于套利空间的最大化，因为在套利策略中，高频数据通常表现得更为出色。另外，由于设备的限制，参数调整存在一定的局限性，因此在实际的金融交易活动中，可以进一步优化和完善策略，以获得更优的结果。

参考文献

[1] Avellaneda M, Lee J H. Statistical arbitrage in the us equities market[J]. Quantitative Finance, 2010, 10 (7):761–782.

[2] Bing-Ting FAN. Futures Arbitrage of Different Varieties and based on the Cointegration Which is under the Framework of Bayesian-In the Case of Soy Oil and Palm Oil[C].E-commerce Development,Shanghai,2017,321-325.

[3] Bogomolov T. Pairs trading based on statistical variability of the spread process[J]. Quantitative Finance, 2013, 13(9): 1411-1430.

[4] Burgess AN.Statistical arbitrage models of the FTSE 100[J]. International Conference Computational Finance, 1999,10:297-312.

[5] CL Dunis, Jason Laws, Ben Evans. Trading futures spread portfolios: applications of higher order and recurrent networks[J]. The European Journal of Finance, 2008, 14(6).

[6] CL Dunis, Jason Laws, Peter W. Middleton,Andreas Karathanasopoulos. Trading and hedging the corn/ethanol crush spread using time-varying leverage andnonlinear models[J]. The European Journal of Finance,2015,21(4).

[7] Daiki Matsunaga, Toyotaro Suzumura, Toshihiro Takahashi. Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis[J].CoRR,2019,abs/1909.10660.

[8] Do B, Faff R. Are pairs trading profits robust to trading costs?[J]. Journal of Financial Research, 2012, 35(2): 261-287.

[9] Dunis C L, Laws J, Evans B. Trading futures spreads: an application of correlation and threshold filters[J]. Applied Financial Economics, 2006, 16(12):903-914.

[10] Elliott R J, Hoekd V D, Malcolm W P. Pairs Trading[J]. Quantitative Finance, 2005, 3(5):271-276.

[11] Engle R F, Granger C W J. Co-Integration and Error Correction: Representation, Estimation, and Testing[J]. Econometrica, 1987,55(2):251-276.

[12] Fischer T , Krauss C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research,2017,

270(2).

[13] Ganapathy Vidyamurthy. Pairs Trading quantitative methods and analysis[M].John wiley&Sons, 2004.

[14] Gatev E, Goetzmann W N, Rouwenhorst K G. Pairs Trading: Performance of a Relative-Value Arbitrage Rule[J]. Review of Financial Studies, 2006, 19(3):797-827.

[15] Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning[J]. The Review of Financial Studies, 2020, Vol. 33, No. 5, PP 2223-2273.

[16] Guijarro-Ordonez J., Pelger M., Zanotti G.Deep Learning Statistical Arbitrage, 2021, Available at SSRN 3862004.

[17] Huang Chien-Feng,Hsu Chi-Jen,Chen Chi-Chung,Chang Bao Rong,Li Chen-An. An Intelligent Model for Pairs Trading Using Genetic Algorithms[J].Computational intelligenceand neuroscience,2015,2015.

[18] Kamalov F. Forecasting significant stock price changes using neural networks[J].Neural Computing and Applications, 2020, 32(1).

[19] Kim SangHo and Park DeogYeong and Lee KiHoon.Hybrid Deep Reinforc ement Learning for Pairs Trading[J].Applied Sciences,2022,12(3):944-944.

[20] Li Lin, Longbing Cao. Mining in-depth patterns in stock market[J]. Int. J. of Intelligent Systems Technologies and Applications,2008,4(3/4).

[21] LI X, WANG J, JIA H, et al. Stock market volatility prediction method based on graph neural network with multi-attention mechanism[J]. Journal of Computer Applications,2022,42(7): 2265.

[22] ML Li, CM Chui, CQ Li. Is pairs trading profitable on China AH-share markets?[J].Applied Economics Letters,2014,21(16).

[23] Montana G, Parrella F. Data mining for algorithmic asset management[M].Data mining for business applications. Springer, Boston, MA, 2009: 283-295.

[24] Mulvey J. M., Sun Y., Wang M., Ye J. Optimizing a portfolio of mean-reverting assets with transaction costs via a feedforward neural network[J]. Quantitative Finance, 2020,Vol. 20, No.8, PP 1239-1261.

[25] Nikos S Thomaidis. Efficient Statistical Analysis of Financial Time-Series Using

Neural Networks and GARCH Model[J]. Working Paper, 2006.

[26] Qiu J,Wang B,Zhou C.Forecasting stock prices with long-short term memory neural network based on attention mechanism[J].PloS one,2020,15(1):e0227222.

[27] Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, Jaewoo Kang. HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction[J]. CoRR,2019,abs/1908.07999.

[28] Rudy J, Dunis C, Giorgioni Getal.Statistical Arbitrage and High-Frequency Data with an Application to Eurostoxx 50 Equities[J].SSRN Electronic Journal,2010.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[30] Wang Q., Teng B., Hao Q., Shi Y., 2021, High-frequency statistical arbitrage strategy based on stationarized order flow imbalance[J]. Procedia Computer Science, Vol. 187, PP 518-523.

[31] Xu Z, Zhang J, Wang J, et al. Prediction research of financial time series based on deep learning[J]. Soft Computing, 2020, 24(6).

[32] 安云博.关于协整的统计套利[J].时代金融,2013.

[33] 蔡燕，王林，许莉莉.基于随机价差法的配对交易研究[J].金融理论与实践,2012(08):30-35.

[34] 陈卫华,徐国祥.基于深度学习和股票论坛数据的股市波动率预测精度研究[J].管理世界,2018, 34(1): 180-181.

[35] 邓晓卫,章铖斌.基于混合神经网络模型预测下的统计套利研究[J].统计与决策,2019,1.

[36] 丁秀玲,华仁海.大连商品交易所大豆与豆粕期货价格内部及之间的套利研究[J].统计研究,2007,24(02): 55-59.

[37] 葛翔宇,吴洋,周艳丽.门限协整套利:理论与实证研究[J]. 统计研究,2012,03:79-87.

[38] 何至静,金苏丹. 基于协整理论的钢铁行业配对交易策略实现 [J]. 投资与创业, 2020, 31 (23): 14-16.

[39] 胡伦超,余乐安,汤铃.融资融券背景下证券配对交易策略研究——基于协整和

距离的两阶段方法[J].中国管理科学,2016,24(4):1-9.

[40] 胡文伟,胡建强,李湛,周剑峰.基于强化学习算法的自适应配对交易模型[J].管理科学,2017,30(02):148-160.

[41] 华仁海,仲伟俊.对上海期货交易所金属铜量价关系的实证分析[J].统计研究,2002.

[42] 黄晓薇,余湄,皮道羿.基于 O-U 过程的配对交易与市场效率研究[J].管理评论,2015,27(01):3-11.

[43] 李斌,邵新月,李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济,2019.

[44] 李世伟.基于协整理论的沪深 300 股指期货跨期套利研究[J].中国计量学院学报,2011,22(02):198-202.

[45] 李晓寒,贾华丁,程雪,李太勇.基于改进遗传算法和图神经网络的股市波动预测方法[J].计算机应用,2022,42(05):1624-1633.

[46] 刘阳,李艳丽,陆贵斌.基于信息更新 NN-GARCH 模型的统计套利研究[J].统计与决策,2016.

[47] 欧阳红兵,李进.基于协整技术配对交易策略的最优阈值研究[J].投资研究,2015,34(11):79-90.

[48] 吴振翔, 陈敏. 中国股票市场弱有效性的统计套利检验[J]. 系统工程理论与实践, 2007,27(2): 92-92.

[49] 邢知,郝继升.基于协整-GARCH模型的统计套利策略最优阈值改进研究[J].延安大学学报:自然科学版,2018,37(03):41-45.

[50] 杨青,王晨蔚.基于深度学习 LSTM 神经网络的全球股票指数预测研究[J].统计研究,2019,36(3): 65-77.

[51] 杨云飞,鲍玉昆,胡忠义,张瑞.基于 EMD 和 SVMs 的原油价格预测方法[J].管理学报,2010, 7(12):1884-1889

[52] 于玮婷.基于协整方法的统计套利策略的实证分析[J].科学决策,2011,03:70-85.

[53] 袁晨,傅强.我国股指期现货的动态相关性及其套期保值效果:来自上证 50、沪深 300 和中证 500 指数的新证据[J].系统工程,2017,35(10):13-22.

[54] 张河生,闻岳春.基于参数调整的协整配对交易策略:理论模型及应用[J].西部

金融,2013.

[55] 张戡,李婷,李凌飞.基于聚类分析与协整检验的 A 股市场统计套利策略[J].统计与决策,2012(15):166-169.

[56] 张鹰. 统计套利理论研究以及匹配股票的协整分析 [J]. 商业经济, 2023, (10): 170-174+193. DOI:10.19905/j.cnki.syjj1982.2023.10.054

[57] 赵胜民,闫红蕾.A 股市场统计套利风险实证分析[J].管理科学,2015,28(5):93-105.

[58] 周亮,陈辰,李宁.基于机器学习和经验模态分解的跨期套利研究[J].西南大学学报(自然科学版),2022,44(01).

[59] 朱丽蓉,苏辛,周勇.基于我国期货市场的跨期套利研究[J].运筹与管理,2015,24(03):179-188.

致 谢

行文至此，落笔为终。在打出“致谢”二字时，代表着我二十余年的求学生涯至此结束。目光所及，皆是回忆，学生胡娜将于 2024 年夏，告别学生时代，结束我炽热的青春。万般不舍，心存感激，好在，轻舟已过万重山。

桃李不言，下自成蹊。首先要感谢我的导师韩海波老师，学生朽木，论文的逐步成型，离不开老师的悉心指导和严谨教诲。得遇良师实乃人生幸事。感谢统计学院各位授课老师及研秘老师，授我学业知识更于人生道路上予我提点。感谢答辩委员会的各位专家，给予我对此论文的深刻指导，百忙中与我一同见证我青春的句号。

焉得谖草，言树之背。最为感谢我的父母、姐姐，父母将我带来这个世界，又有姐姐陪伴我、引导我成长，让我始终站在他们的肩膀上感受这世间的温柔，感谢家人在我成长中给予我爱与包容，感谢家人给予我的底气和力量，感谢家人护我于柔软的怀抱未曾经历苦难。

岁月清浅，时光激滟。回望求学生涯，始于初秋终于夏，感谢遇见的所有同学，也感谢如今陪伴在身边的挚友。

路漫漫其修远兮，吾将上下而求索。感谢一路上走的很慢但始终向前的自己，始终勇敢而真挚。

何其有幸，生于华夏。生逢于盛世，感谢国家，祝我的祖国繁荣昌盛。

追风赶月莫停留，平羌尽处是春山，愿我们前路漫漫亦灿灿。