**目录**

[1 引言 3](#_Toc5328)

[1.1 研究背景及研究意义 3](#_Toc3913)

[1.1.1 研究背景 3](#_Toc29245)

[1.1.2 研究意义 3](#_Toc4983)

[1.2 文献综述 4](#_Toc19963)

[1.2.1 商品期货套利研究进展 4](#_Toc3718)

[1.2.2 套利模型及策略方面的研究 5](#_Toc23697)

[1.3 研究内容 6](#_Toc25505)

[(1) 数据获取与预处理 6](#_Toc23117)

[(2) 相关性与协整性分析 6](#_Toc29794)

[(3) 交易策略设计与实施 6](#_Toc1734)

[(4) 机器学习模型的引入 7](#_Toc23780)

[1.4 本文创新点 7](#_Toc25882)

[1.5 本文整体结构 7](#_Toc21873)

[2 配对交易策略理论基础及具体操作方法 9](#_Toc12147)

[2.1均值回归理论 9](#_Toc3723)

[2.2主成分分析和OPTICS聚类 9](#_Toc8023)

[2.3波动率网络模型 10](#_Toc17000)

[2.4 时间序列平稳性分析 11](#_Toc31178)

[2.4.1平稳序列 11](#_Toc20243)

[2.4.2平稳序列检验 11](#_Toc1369)

[2.5 协整性及协整性检验 12](#_Toc1746)

[2.6 条件异方差模型 13](#_Toc376)

[2.6.1 ARCH模型 13](#_Toc4226)

[2.6.2 GARCH模型 13](#_Toc12767)

[2.7机器学习模型 14](#_Toc16389)

[2.7.1 LSTM模型 14](#_Toc26061)

[2.7.2 Xgboost模型 15](#_Toc20219)

[2.8配对交易策略具体操作方法 15](#_Toc25171)

[2.8.1数据调用及预处理 15](#_Toc24661)

[2.8.2确定交易策略 15](#_Toc4125)

[2.9 策略评估 16](#_Toc19854)

[3 实证数据预处理与协整性分析 18](#_Toc4173)

[3.1数据获取及预处理 18](#_Toc29319)

[3.2高相关性期货对获取 19](#_Toc14694)

[3.2.1主成分+OPTICS聚类 19](#_Toc5460)

[3.2.2网络模型搭建 20](#_Toc17497)

[3.3协整性分析 21](#_Toc12681)

[4 实证策略构建 23](#_Toc29490)

[4.1实证策略构建 23](#_Toc27324)

[4.2策略参数优化 23](#_Toc1530)

[4.3时变标准差模型引入 24](#_Toc18528)

[4.3.1 ARIMA模型拟合 24](#_Toc14271)

[4.3.2策略制定及策略参数优化 26](#_Toc11552)

[4.4基于基差的策略配置 27](#_Toc3311)

[4.4.1基差研究 27](#_Toc20065)

[4.4.2策略配置及参数优化 28](#_Toc17758)

[5 拓展策略构建 30](#_Toc21660)

[5.1机器学习效果 30](#_Toc6068)

[5.1.1 LSTM神经网络 30](#_Toc21412)

[5.1.2 XGboost 31](#_Toc18713)

[5.2机器学习效果实证 31](#_Toc543)

[6 总结与展望 33](#_Toc2892)

[6.1 总结 33](#_Toc4534)

[6.2 不足与展望 34](#_Toc7320)

[(1) 交易成本的考虑不足 34](#_Toc21954)

[(2) 特征变量的扩展 34](#_Toc16446)

[(3) 机器学习模型的进一步优化 34](#_Toc5262)

**摘要**

配对交易作为中国期货市场中的重要统计套利手段，凭借其市场中性特征和稳定的交易逻辑，已成为量化投资的重要组成部分。本文基于配对交易的基本逻辑，深入探讨其在中国大宗商品期货市场中的应用，重点研究强相关性期货品种的套利机会。本文通过协整分析和条件异方差模型（GARCH模型）优化策略，并结合机器学习模型（LSTM与XGBoost）探索其在期货市场中的实际表现。

本文首先介绍了配对交易的理论基础，并采用主成分分析（PCA）和OPTICS聚类算法筛选出波动性高度相关的期货对。随后通过ADF单位根检验与协整分析验证菜油和棕榈油期货之间的长期均衡关系。针对该期货对，本文构建了价差交易策略，以去中心化价差为核心设定交易阈值，并制定详细的开仓、平仓与止损规则。在基本价差策略的基础上，引入对冲比率优化价差计算，并通过超参数优化寻找到最优的阈值参数。

此外，本文引入GARCH模型计算时变标准差，替换原有的静态标准差，以实现策略的动态调整和更精确的风险控制。为进一步提升收益，本文还使用LSTM与XGBoost模型预测未来价差序列，优化短期与长期交易决策。实证分析表明，配对交易策略在大宗商品期货市场中表现出色，特别是XGBoost模型实现了较高的年化收益率和较小的回撤，而LSTM模型则在长期趋势捕捉上表现稳定。本文的研究展示了配对交易策略在大宗商品市场中的广泛应用潜力，并为投资者提供了系统化的套利交易参考。

**关键词：**配对交易 协整分析 条件异方差模型 LSTM XGBoost

**ABSTRACT**

Pair trading, as an essential statistical arbitrage strategy in China's futures market, has become a vital component of quantitative investment due to its market-neutral characteristics and stable trading logic. This paper explores the application of pair trading in China’s commodity futures market, focusing on arbitrage opportunities among highly correlated futures contracts. The study optimizes trading strategies through cointegration analysis and the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, and further incorporates machine learning models (LSTM and XGBoost) to evaluate the strategy’s performance in the futures market.

The paper first introduces the theoretical foundation of pair trading and applies Principal Component Analysis (PCA) and OPTICS clustering to identify futures pairs with strong volatility correlations. ADF unit root tests and cointegration analysis are then performed to validate the long-term equilibrium relationship between rapeseed oil and palm oil futures. Based on the selected futures pair, the study constructs a spread trading strategy, using detrended price spreads to set trading thresholds, and establishes detailed opening, closing, and stop-loss rules. The basic spread strategy is further enhanced by introducing a hedging ratio for spread calculation, and hyperparameter optimization is employed to identify the optimal trading thresholds.

In addition, the GARCH model is used to calculate time-varying standard deviations, replacing static standard deviations to enable dynamic threshold adjustments and more precise risk management. To further enhance returns, the study applies LSTM and XGBoost models to predict future spread movements, optimizing both short-term and long-term trading decisions. The empirical results demonstrate that pair trading strategies perform well in the commodity futures market. Notably, the XGBoost model achieved a high annualized return with low drawdown, while the LSTM model showed robust performance in capturing long-term trends. This research highlights the extensive application potential of pair trading strategies in the commodity market and offers systematic arbitrage trading guidance for investors.

**Keywords**: Pair trading, Cointegration analysis, GARCH model, LSTM, XGBoost

# 1 引言

## 1.1 研究背景及研究意义

### 1.1.1 研究背景

配对交易策略最早由高盛的研究员Nunzio Tartaglia提出，作为一种量化投资手段。他通过分析一系列具有关联性的资产，发现其中隐藏的相互关系，并将其应用于真实交易中。配对交易的核心思路是利用两种或多种高度相关的资产，通过观察其价格变化之间的差异来寻求套利机会。该策略的理论基础主要包括协整关系和统计套利两部分。协整关系认为，即便两种资产的价格表现看似独立，它们之间可能仍然存在一种长期且稳定的联系，可以通过统计方法来识别和量化。一旦确认这种关系存在，当市场价格偏离其长期均衡水平时，可以预期价格最终会回归均衡，从而实现利润。

与协整关系不同，统计套利更关注那些资产之间的短期价格联动。这种短期的相关性通常来源于市场情绪、技术信号或基本面因素的变化，为交易者提供了捕捉暂时性价格偏离的机会。基于这种理论，投资者可以通过挑选具有相关性的资产组合，并使用统计方法分析其价格关系，以此确定最佳交易时机。

由于该策略建立在严谨的统计分析基础之上，能够有效捕捉市场价格波动的特征，因此近年来受到越来越多投资者的青睐。配对交易具有市场中性的特征，即无论市场处于何种环境下，其交易逻辑和收益模式都能保持一定的稳定性。这一优势吸引了许多投资者尝试使用该策略，以期在多样化的市场条件下获得持续的收益。

随着中国期货市场的快速发展，大宗商品期货逐渐成为投资者关注的重点。自1990年中国引入期货交易以来，市场规模不断扩大，量化投资逐步成为主流投资方式之一。在这一趋势下，配对交易作为一种重要的量化策略，不仅为投资者提供了新的交易思路，也为期货市场注入了更多活力。随着中国资本市场的开放程度不断提高，国内投资者得以参与全球市场，通过跨境配对交易策略实现更广泛的资产配置，跨市场交易也逐渐成为资产管理的重要手段。

### 1.1.2 研究意义

在期货交易中，统计套利作为一种重要的量化投资策略，具有广泛的应用前景。然而，如何将统计套利策略有效地应用于中国期货市场，仍然存在一定的挑战。随着中国资本市场快速发展，量化投资逐步成为主流投资方式之一。配对交易作为统计套利的重要组成部分，因其市场中性特征在不确定性市场环境下表现出较好的稳定性。本研究结合中国市场环境，开发了适应本地市场特性的套利策略，为投资者提供了一种行之有效的量化交易工具。

本研究努力在传统的金融模型基础上，引入了GARCH模型进行时变风险的测量与控制，并结合LSTM（长短期记忆网络）和XGBoost（梯度提升树）等机器学习算法。通过这些技术，增强了模型的预测能力，优化了短期和长期交易决策。这种融合策略为投资者在应对市场波动时提供了灵活且高效的解决方案。

随着中国资本市场开放程度的提高，投资者在国内市场的交易越来越多地与全球市场接轨。本研究的配对交易策略，不仅可以在中国市场应用，还可通过跨市场套利策略参与国际市场。这为国内投资者的全球资产配置提供了理论支持和实践参考。在交易过程中，本研究不仅通过协整分析确保选定期货品种间的长期均衡关系，还通过GARCH模型的时变标准差优化开仓和平仓阈值，确保策略能有效应对市场波动。此外，研究引入了超参数优化，进一步提升了策略的收益稳定性，为投资者降低了潜在风险。

## 1.2 文献综述

### 1.2.1 商品期货套利研究进展

自我国1990年进入期货交易领域以来，期货市场经历了快速发展。量化投资作为一种重要的市场策略，通过建立数学模型和计算机程序，实现了交易的自动化操作。特别是在统计套利领域，量化投资取得了显著的成果。其中，配对交易作为一种经典的统计套利策略，已在国内外市场中得到了广泛的研究与应用。

国内关于期货套利的研究起源于钱小安对套利交易的探索。他的研究基于对两个相同或类似金融工具之间价格矛盾的观察和预测，试图利用图形分析或数学模型来量化价格异常的程度。这些研究认为，历史数据会影响交易者的预期，从而促使他们采取不同的交易策略。在当时，这种方法被广泛应用于国债套利领域。此外，毛二万和宋逢明构建了一个包含未定债权交易和投资套利的双期经济框架，并提出了套利机会的定义——即通过某些未定债权的组合，在不确定的市场条件下也能获得收益。他们的研究表明，套利不仅能为价格异常提供信号，还能通过这些信号获取交易利润。

在国际研究方面，Leo Melamed 指出，配对交易的部分收益可能源于市场结构性因素，如非同步交易和买卖价差的波动。他们测试了多种自筹资金的交易策略，发现顶级配对组合的年化超额收益率约为12%。即便考虑到交易成本，这些策略依然能够实现可观的正收益。此外，Kanamura、Rachev 和 Fabozzi 提出了一个针对价差交易的利润模型，重点分析价差的随机波动及其首次达到某水平的概率密度。他们还修改了模型以适应能源期货市场，发现可以用均值回复过程来描述能源期货价差的变化。由于该过程的首次到达时间的概率密度可以较为准确地估计，使用这些参数即可构建能源期货价差的利润模型。David P. Simon 的研究表明，大豆压榨价差具有明显的季节性和均值回归特性，这些特性可用于制定有效的交易策略。

### 1.2.2 套利模型及策略方面的研究

在金融市场中，套利交易策略扮演着至关重要的角色，其核心在于利用不同市场或合约之间的价格差异来实现盈利。这类策略涵盖了经典套利理论、统计套利以及机器学习在套利中的应用等多个方面。

经典套利理论的基础是无风险套利的概念，即通过捕捉市场价格差异，在不承担任何风险的情况下获得利润。Ross 提出的套利定价理论（APT）是该领域的重要代表之一。APT 假设市场价格受到多个风险因素的共同影响，并通过构建无风险套利组合，使市场价格回归均衡，实现合理定价。

统计套利是一种基于历史数据与统计模型的套利策略，广泛应用于配对交易中。Gatev、Goetzmann 和 Rouwenhorst 的研究表明，配对交易可以通过选择历史价格走势相关的股票对来进行。当两只股票价格出现背离时，买入价格较低的股票、卖出价格较高的股票，并在两者价格回归时获利。Engle 和 Granger 提出的协整模型则提供了理论支持，该模型指出，即便两种资产价格表面上无相关性，它们之间仍可能存在长期稳定的协整关系，从而为统计套利提供了操作依据。Laila、Rahman 和 Saleem 进一步分析了协整模型在卡拉奇证券交易所中的应用，研究如何通过协整关系筛选适合进行统计套利的股票对。Ahmet 和 Erdinç 则从实证角度评估了不同配对交易策略在统计套利中的表现，验证了这些策略的有效性。

随着技术的进步，机器学习和人工智能逐渐被引入套利交易中，进一步提升了策略的收益与稳定性。Avellaneda 和 Lee 提出了一种基于机器学习的算法交易策略，利用模型预测资产的价格走势，优化交易策略，并提升风险调整后的回报。他们采用了回归分析、主成分分析和克里金插值法等技术，以捕捉股票对之间的短期价格偏离，并通过实证分析证明了该策略在市场套利中的有效性。Krauss、Do 和 Huck 的研究显示，与传统统计方法相比，机器学习算法在配对交易中的表现更为优越。他们重点分析了价差的随机运动及其首次到达时间的概率分布，并将这一模型应用于能源期货市场。通过均值回复过程对能源期货价差进行建模，发现基于首次到达时间概率密度的套利模型在能源期货市场中表现出良好的适用性。Lam 则探索了机器学习与多区域纠缠熵相结合的统计模型，展示了如何利用机器学习技术捕捉复杂金融市场中的套利机会，并证明了这些算法在提升预测精度和优化交易决策方面的有效性。

## 1.3 研究内容

本文围绕大宗商品期货市场中的配对交易策略展开研究，聚焦于多个商品期货对的关联性和套利机会，结合协整分析、GARCH模型以及机器学习技术，探讨多种策略的构建与优化。研究内容涵盖以下几个方面：

(1) 数据获取与预处理

从Wind数据库选取2018年1月1日至2023年12月31日期间的主要商品期货合约，获取期货合约的基本行情数据。针对缺失值采用前向填充或后向填充法进行补全，并剔除异常数据，确保数据的完整性和准确性。针对不同商品的价格规模差异，对数据进行标准化和归一化，为后续建模提供一致性数据格式。

(2) 相关性与协整性分析

计算不同商品期货波动率的相关系数矩阵，利用网络模型筛选出具有高相关性的期货对，初步锁定具有套利潜力的组合。使用ADF检验对期货价格序列的平稳性进行验证，并进一步通过协整分析判断其是否存在长期均衡关系。协整性较强的期货对作为配对交易的核心标的，以降低虚假回归的风险。

(3) 交易策略设计与实施

根据协整性分析结果，计算去中心化后的价差序列，并基于价差的均值和标准差设定开仓与平仓阈值。当价差序列的绝对值超过设定阈值时开仓，回归到阈值内时平仓。引入GARCH模型捕捉价差序列的时变波动性，根据市场的动态变化调整开仓和平仓阈值，以实现更好的风险控制和收益优化。针对期货与现货市场之间的差异构建基差模型，研究不同市场条件下基差的回归特性，以此作为交易策略的补充手段。

(4) 机器学习模型的引入

设计LSTM模型预测未来价差的走势，捕捉长期趋势，用于指导配对交易的开仓与平仓决策。基于XGBoost算法训练回归模型，快速预测短期市场波动，帮助优化短期高频交易策略。通过两种机器学习模型的比较，探索长期与短期市场波动的不同应对策略。

## 1.4 本文创新点

和过往的研究内容相比，本文的创新点有：

（1）创新型主成分分析和网络模型构建：本文创新性地将主成分分析（PCA）与OPTICS聚类相结合，从多品种期货的波动率数据中提取主要特征，减少冗余信息，提高分析效率。在此基础上，构建了基于波动率相关系数的网络模型，将期货品种作为节点，通过其波动率相关性建立连接。该网络模型直观展示了品种之间的动态关系，识别出如铜与铝、豆粕与菜粕等高度相关的集群，为筛选配对交易组合提供科学依据。这一方法提升了配对交易品种选择的精准性和可视化效果，超越了传统相关性分析的局限。

（2）多模型融合应用：本文不仅使用了协整分析和GARCH模型等传统金融模型，还创新性地引入了LSTM和XGBoost机器学习算法，构建不同时间跨度的交易策略。通过融合短期与长期模型的优势，XGBoost模型在捕捉短期波动方面表现出色，而LSTM模型在长期趋势追踪上具有更好的稳定性，提升了整体策略的灵活性和表现。

（3）多维度数据处理与基差模型的引入：除了常见的价差交易策略，本文还构建了基差模型，分析期货与现货价格之间的关系，拓展了传统配对交易策略的应用范围。基差模型使交易者能够更好地捕捉不同市场间的价差变化，从而提供新的套利机会。

## 1.5 本文整体结构

本文将分为六个章节展开，对大宗商品期货市场中的配对交易策略进行全面的分析和探讨。第一章为文章的背景介绍与研究意义，并结合已有文献对配对交易策略的发展与应用进行回顾。该部分将详细阐述配对交易在金融市场中的作用，重点探讨其在期货市场中的应用价值与挑战，明确本研究的核心问题和创新点。

第二章介绍配对交易的理论基础与分析方法，讨论时间序列分析中的平稳性与单位根检验（如ADF检验）。本章还深入介绍了协整性分析及其在识别资产价格长期均衡关系中的应用，特别是E-G两步法的使用。此外，条件异方差模型和机器学习模型也在本章中进行了阐述，解释其在捕捉金融时间序列波动性方面的重要性。第三章主要描述数据的获取与预处理过程。本文选取多个大宗商品期货的价格数据作为研究对象，详细介绍了数据清洗、缺失值处理和标准化的步骤，确保数据质量。随后，通过协整分析与相关性检验对数据进行实证分析，判断其是否适合用于配对交易，并使用网络模型筛选出具有套利潜力的期货对。第四章和第五章详细介绍配对交易策略的构建与回测。本章讨论了如何构建去中心化后的价差序列，并基于均值和标准差设置交易阈值，决定开仓与平仓的具体条件。除了基本的对冲价差模型，还通过引入GARCH模型、基差模型、LSTM和XGBoost模型构建不同的策略组合，并对每个策略进行回测分析，比较其在不同市场条件和参数组合下的表现。第六章对研究进行总结与展望。本章将系统比较各策略的收益、风险和适用性，分析不同策略在捕捉短期与长期市场机会方面的优势与不足。最后提出未来研究方向，包括交易成本控制、跨市场配对交易及其他品种的探索，并讨论如何进一步优化现有模型，以提升策略的稳定性与收益水平。

# 2 配对交易策略理论基础及具体操作方法

## 2.1均值回归理论

在统计学领域，均值回归理论是建立在正态分布假设的基础之上的，它主张在长期的演变过程中，事物会倾向于回到其均衡点。这一理论最初由FrancisGalton通过广泛的实验揭示，特别是在他对甜豌豆种子的研究中，他注意到子代豌豆的直径分布相较于母代更为集中。该现象在自然界和社会现象中均有所体现，并推动了多种风险评估和预测理论的形成。均值回归理论在期货交易中也占据着举足轻重的地位，被视为金融理论的一个关键组成部分。在应用此理论时，需要首先计算出期货价格在特定时间段内的平均数值。依据均值回归的观点，无论价格如何波动，无论是高于还是低于预期的均值，它最终都将趋向于这个均值。

因此，当期货价格在一段时间内表现出显著的上升或下降趋势，明显偏离其平均值时，未来的价格走势有很大概率会向均值靠拢。对于投资者来说，这种价格的偏离实际上提供了一个理想的交易契机。但在实际操作中，由于众多投资者都会迅速注意到这种价格偏离，因此交易机会的窗口往往比较短暂，价格也会因此迅速得到调整。所以，投资者必须依赖精确的统计技术分析来及时捕捉交易信号。

在进行期货配对交易时，选择的两份期货合同应该具有长期的均衡关系，这通常意味着它们之间的价差在长时间内保持相对稳定。然而，当某个外部因素导致其中一份合同的价格出现大幅波动时，两者之间的价差就会偏离其常规范围，原有的平衡状态就会被打破，一份合同的价格会显得相对较高，而另一份则相对较低。在这种情况下，策略是购买价格偏低的合同，并出售价格偏高的合同。根据均值回归的原理，这种价格失衡的状态不会持续太久，当价格回归到正常水平——即两者的价差恢复稳定时，再进行相反的操作。这就是配对交易的基本流程，通过这种方式可以实现利润的增长。

## 2.2主成分分析和OPTICS聚类

主成分分析是一种常用的降维技术，旨在通过线性变换将高维数据投影到一个较低维的空间中，从而减少数据的维度，同时尽可能保留原始数据中的主要信息。在金融市场数据中，PCA常用于分析多维时间序列数据的主要变化趋势。通过计算标准化后数据的协方差矩阵，度量不同变量之间的相关性，提取协方差矩阵的特征向量（主成分方向）和特征值（各主成分的方差大小）。根据累计解释方差的贡献率，选择解释方差比例较高的前几个主成分，作为数据的新维度表示。本文计算每日价格波动率并将每日价格波动率视作高维变量，使用主成分分析降维至若干代表性日期波动率以衡量该品种在长周期价格波动的表现。

OPTICS是一种基于密度的聚类算法，专为处理具有不同密度的复杂数据结构设计。与其他密度聚类算法（如DBSCAN）相比，OPTICS能够识别不同密度的簇，而无需预先指定明确的密度阈值或簇的数量。OPTICS的核心概念包括核心距离和可达距离：核心距离表示某个数据点成为簇核心所需的最小邻域半径，即其周围必须至少有指定数量的点，而可达距离定义为某个数据点到其邻域中其他点的最小可达距离，用来判断数据点之间的聚类关系。OPTICS的关键步骤是根据核心距离和可达距离，对数据点进行有序排列，生成“可达性图”，从中可以动态识别不同密度的簇并自动发现噪声点。OPTICS特别适合不规则形状和密度变化较大的数据集，无需事先设定簇数，能够灵活处理密集与稀疏区域并识别出嵌套的簇结构。本文使用降维后的日价格波动率数据进行OPTICS聚类，最终得到高相关性的商品期货对。

### 2.3波动率网络模型

为识别商品期货之间的关联性，本文采用基于波动率的网络模型构建方法。此方法基于不同商品期货在价格波动上的一致性，通过计算期货品种间的波动相关系数，来判断哪些商品对适合用于配对交易。根据龙奥明的研究，日频数据下的波动率计算能够较为准确地反映商品期货的波动特征。因此，本文使用日内最高价、最低价与收盘价数据，计算每个品种的每日波动率。波动率的计算公式如下：

其中，表示第n个商品期货在某个交易日的收盘价，最高价和最低价分别表示该商品在该交易日的价格区间。在得到各品种的波动率后，本文进一步计算任意两个商品期货间的波动率相关系数，公式如下：

其中为两个商品i和j之间的波动率相关系数，和分别表示商品i和j的波动率，和分别表示他们的波动率标准差。通过计算每个商品期货之间的波动率相关系数构建一个无向加权网络，在这个网络中，每个商品期货被视为一个节点，节点之间的边权重由波动率相关系数决定。波动率相关系数较大的商品对在波动性上具有较强的联系，因此可以作为潜在的配对交易对象。

## 2.4 时间序列平稳性分析

### 2.4.1平稳序列

在时间序列分析里，平稳性是基本的一个概念。平稳性意味着随机项会沿着水平值波动，并且前后之间具有相关性，与独立序列不同。如果时间序列满足：

（1）对任何；

（2）对任何；

（3）对任何。

那么称是平稳的时间序列，简称为平稳序列。称实数为的自协方差函数。

本文处理的期货量价数据为时间序列数据，这与传统统计数据的格式有着明显的不同。主要差异在于，传统的统计数据常可对同一随机变量进行重复观测，但在时间序列数据中，每一时间节点上的随机变量仅对应一个观测值。本文所涉及的期货收盘价数据，每个交易日仅有一个收盘价且不同日期的价格分布各异。因此，时间对于序列的影响必须被考虑。然而，每一时间点的观测值仅代表该时间点的特定分布，无法通过重复观测同一时间点来深入探究其分布特性。这种数据构造使得样本量相对较小，进而影响了分布估计的精确度。可以通过探究数据的稳定性来发现不同日期分布之间的内在联系，从而在一定程度上弥补因样本量不足而导致的精度问题，确保了时间序列在不同时间区间内保持一致的统计特性。稳定性也是这种历史数据和未来数据相似性的保障，一个稳定的时间序列意味着其历史特性有可能在未来重现，这为利用历史数据进行统计套利分析预测提供了依据。

### 2.4.2平稳序列检验

学界有关于时间序列平稳性的检验更多的是单位根检验，具体而言为DF检验、ADF检验，其大致思路是判断一个时间序列的特征根在单位圆内还是单位圆外，并依据此判断序列的平稳性。

DF检验由Dickey和Fuller发明，该方法的前提为假设时间序列是由1阶自回归过程，即AR(1)产生。基于非平稳序列的基本特征，可以将序列大致归为三类，并且对其进行假设检验：

（1）无漂移项自回归过程，特征为序列呈现无规则上升或下降；

（2）带漂移项自回归过程，特征为序列呈现明显的随时间递增或递减且趋势不强；

（3）带趋势项回归过程，特征为序列随时间快速递增时。

其中是常数项，是时间趋势项，为白噪声无自相关性，为AR(1)模型拟合系数。并由此进行假设检验，

。

原假设即为1，意味着存在单位根，时间序列是非平稳的；备择假设为时间序列是平稳的。通过检验统计量的p值是否大于显著性水平α来进行判断。

为了使DF适用于高阶自回归过程，Dickey等人引入了更多的滞后项，从而发展出ADF检验。ADF检验通过增加滞后差分项来解决序列自相关的问题。本文使用ADF检验用于判断时间序列的平稳性。

## 2.5 协整性及协整性检验

在真实的市场之中，由于收到市场事件、政府政策、购买者消费心态等因素的冲击，金融产品的量价数据往往存在很大的波动，金融时间序列通常是不平稳的。即使存在某些产品，例如可以互为替代品的菜油和棕榈油，存在逻辑意义上的关联性，反映在价格序列中即为长期均衡关系，为了防止出现虚假回归问题，不能轻易使用经典线性回归模型。

而对于不平稳的序列，可以通过对序列差分得到平稳的序列。然而，差分的操作将会使得原始序列在时间上的一些特征丢失，这对之后对于序列进行建模会产生影响，因此需要对序列通过其他的定义来判断之间的有效信息的关联性。因此Engle和Granger提出了协整概念。记自变量序列为，因变量序列为，构建回归模型：

。

和分别为该回归模型的参数估计，而为残差项，若回归模型的残差项是平稳的，可认为与存在协整关系。该模型不需要要求序列与平稳。但在证明协整性前，要求它们都是d阶单整即序里至少经过d阶差分才能平稳。

因此对于之前所谈论到的金融数据的非平稳性进行建模，可以先进行协整关系性检验。常用的协整性检验为EG检验。一般情况下，想得到的结果是非平稳序列间具有协整关系，故EG检验的原假设、备择假设分别为

：非平稳序列之间不存在协整关系：非平稳序列之间存在协整关系。

根据之前的结论推导，仅需要对原先的回归方程的残差项进行平稳性判断，即等价于原序列的协整性判断。若通过了EG检验，说明序列间存在协整关系，具有长期均衡关系，此时可以建立经典回归模型来描述均衡关系。

## 2.6 条件异方差模型

有关于配对交易的策略构建主要包含两方面内容，一方面根据之前的均值回复理论，判断价差序列将会返回到之前的均值水平，因此需要通过计算价差序列均值来进行策略构建；另一方面对于价差序列所到达的开仓和平仓阈值，将基于对历史序列的样本进行决定。如果假设价差序列的方差不发生变化，即判断为方差齐性，可用历史序列标准差来作为阈值标准。然而，之前阐述过金融时间序列会受到多方因素冲击而不可直接预测，因此其量价序列所呈现出的标准差也会存在一定的波动，在某些具体的事件冲击下会持续偏大，即不满足方差齐性。因此，很多统计学者更倾向于用异方差模型拟合金融量价序列。

### 2.6.1 ARCH模型

条件异方差模型的基础为ARCH模型，记自变量序列为，因变量序列为该模型基础如下：

。

其中为误差项。如果误差项的平方服从过程，即：

。

其中独立同分布，并满足E。该模型即为自回归条件异方差模型（ARCH）。要求以确保为正值。

ARCH模型主要用于对主体模型中的随机扰动项进行深入分析，旨在全面捕捉残差中的信息，确保最终模型残差符合白噪声序列的标准。由于噪声的方差实际上是过去有限项噪声值平方的函数性表达，即噪声的波动展现了一定的自相关性或记忆性。因此，当过去某一时刻的噪声方差呈现上升趋势时，当前时刻的噪声方差也倾向于增大；反之，若过去某时刻噪声方差减小，当前时刻的噪声方差也倾向于减小。

### 2.6.2 GARCH模型

然而ARCH模型需要估计多个滞后期的参数，当模型阶数较高时，需要估计的参数数量会急剧增加，导致参数估计变得复杂且不稳定。同时，高阶滞后期的参数估计可能导致负方差的问题。为了更好地捕捉金融时间序列表现出的波动聚集情况，Bollerslev建立了广义自回归条件异方差模型。假定干扰序列具有如下结构：

其中，且,t代表时间。该模型则为广义自回归条件异方差模型，记为。本文中选择使用GARCH模型去拟合的价差序列，并且通过得到的异方差作为时变方差去计开仓与平仓的阈值，这样可以更好地去衡量价格市场的波动率变化。

## 2.7机器学习模型

随着金融数据的复杂性增加，传统统计模型在捕捉市场非线性特征方面存在一定的局限性。因此，近年来机器学习模型被广泛应用于金融时间序列预测中。本文结合了LSTM（长短期记忆网络）和XGBoost（极限梯度提升树）两种机器学习模型，用于预测商品期货价差的变化趋势，并进一步用于配对交易策略的构建。

### 2.7.1 LSTM模型

长短期记忆网络是一种递归神经网络的特殊变体，特别适用于处理具有长期依赖性的时间序列数据。传统的RNN容易受到梯度消失或梯度爆炸问题的影响，难以捕捉远距离的依赖关系，而LSTM通过引入记忆单元，能够在较长时间范围内保持重要的时序信息。LSTM的核心在于其特殊的结构：遗忘门、输入门和输出门。这些门控机制能够决定哪些信息应该被记住，哪些应该被遗忘，从而有效解决长时间序列中的依赖性问题。每个LSTM单元由以下公式描述：

（1）遗忘门，决定需要丢弃哪些过去的信息，是权重矩阵，是上一个时刻的隐状态，是当前输入，是偏置项，为Sigmoid激活函数；

（2）输入门，决定更新哪些新的信息，是候选记忆单元；

（3）记忆单元更新：；

（4）输出门：。

LSTM通过上述机制有效捕捉了时间序列中的长短期依赖，在本文中被用于预测期货价差序列的未来走势，进而为配对交易策略提供依据。

### 2.7.2 Xgboost模型

XGBoost是一种集成学习算法，基于梯度提升思想构建，广泛用于结构化数据的建模。XGBoost通过逐步构建多个弱学习器（如决策树）来减少预测误差，并通过最小化损失函数和引入正则化项，来提高模型的泛化能力。XGBoost在每一轮迭代中通过构建新的决策树来修正前一轮模型的误差。其损失函数不仅包含预测误差，还引入了正则化项来防止模型过拟合。损失函数的形式为：

其中，是预测值和真实值之间的损失，是正则化项，用于控制模型复杂度，防止过拟合。XGBoost构建的每一棵树都是为了最小化损失函数中的误差部分。通过加权的方式，每棵树依次修正之前的预测误差。在每次迭代中，XGBoost使用梯度下降法来优化每棵树的预测效果，使得新树能够最大程度上减少前一轮的残差。通过引入正则化项，XGBoost在保证模型拟合效果的同时，控制了模型的复杂性，从而提升了模型的泛化能力。

## 2.8配对交易策略具体操作方法

基于前面所介绍的配对交易理论，下面将介绍本文具体进行配对交易策略构建的操作步骤。

### 2.8.1数据调用及预处理

将选取全行业商品期货的基本市场行情数据，数据清洗后使用主成分分析和OPTICS聚类，结合网络模型，找出符合配对交易假设前提的期货对。进一步判断它们的价格是否具有协整关系，从而进行配对交易策略的构建。

### 2.8.2确定交易策略

对两类期货的价格序列{进行回归，得到回归表达式：

。

其中为回归拟合参数，为残差序列，t为收盘价所处时间。在对冲配对交易中，用系数作为两类期货的对冲比例，即当买入1头寸的期货品种时，就要卖出对应的头寸期货，反之亦然。由此得到

。

对于开仓和平仓的条件判断，基于对与的比较进行，为人为设置的交易阈值，为标准差或异方差。优先讨论标准差的情况，其计算方法分为两种，一种是假定方差齐性，使用样本内的样本标准差作为；另一种是计算时变标准差，即在当前交易日向前推导一个窗口（windows）的时间，以该时间段内的数据作为样本值计算标准差，以此代替。使用标准差的倍作为交易阈值，当大于交易阈值时，判断为存在交易位点。具体规则如下：

（1）当，开仓，卖出价位高的期货标的，买入价格低的期货标的，当时，反向平仓；

（2）当，开仓，卖出价位高的期货标的，买入价格低的期货标的，当时，反向平仓。

使用超参数优化的方法寻找得到能产生最优收益率的𝑤。理论上当价差序列突破所拟定阈值时，根据均值回复理论，会回到均值附近。然而由于多方事件等因素冲击，金融序列可能存在一定的回落延后期，若此时不平仓，投资者将会蒙受很大的沉没成本损失。因此选择定额止损法，通过设定一个亏损上限，当亏损达到或超过这个上限时，自动触发平仓操作，以防止进一步的损失。

## 2.9 策略评估

根据既定策略，在样本期间内，根据开仓和平仓的阈值判断，会进行多次配对交易。暂不考虑交易成本，每次交易后计算出单次交易的收益率公式为：

。

其中，、分别代表开仓、平仓时卖出金额，、分别代表开仓、平仓时买入金额。为了评估策略在样本期内的表现，本文采用年化收益率、最大回撤和夏普比率进行回测分析。年化收益率用于衡量策略的长期收益表现，其公式为：

k表示交易的总次数，T是样本期的总交易天数，252是股票市场的年交易天数。最大回撤用于衡量策略在回测期间的风险，表示回测期间的净值从最高点到最低点的最大跌幅。夏普比率用于衡量策略的风险调整后收益，其公式为：

其中，为策略的收益率，为无风险收益率，为策略收益率的标准差。夏普比率反映了每单位风险所获得的超额收益，值越高表明策略的风险调整收益越好。

# 3 实证数据预处理与协整性分析

## 3.1数据获取及预处理

本次分析中，本文选取了三大期货交易所的所有期货主力合约数据，研究期间为2018年1月1日至2023年12月31日。从wind上获取研究期间内的三大所商品期货主力连续合约对应的基本行情数据：

**表3.1商品期货市场行情数据变量名称**

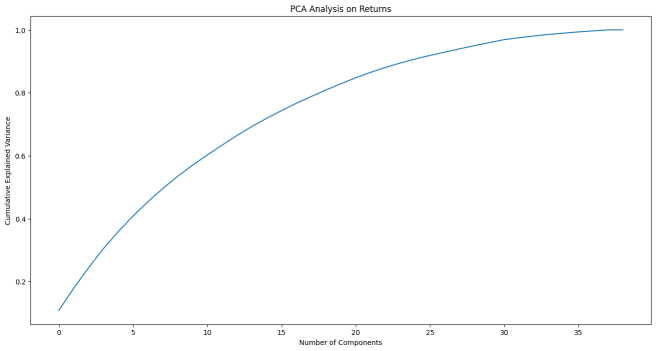
|  |  |  |
| --- | --- | --- |
| 名称 | 类型 | 描述 |
| code | str | 合约代码 |
| close | float | 收盘价 |
| high | float | 最高价 |
| low | float | 最低价 |
| volume | int | 成交量 |
| amt | float | 交易金额 |
| pct\_chg | float | 日涨跌幅（百分比变化） |
| swing | float | 振幅 |
| oi | int | 持仓量 |
| anal\_basis | float | 基差（现货与期货价格的差异） |
| basisannualyield | float | 基差年化收益率 |
| anal\_basispercent2 | float | 基差百分比（以第二腿计算的比例） |
| ftdate | str | 交易日期 |
| lasttrade\_date | str | 最后交易日 |
| contractmultiplier | int | 合约乘数（每手合约代表标的数量） |
| trade\_hiscode | str | 历史合约代码（用于跟踪合约变化） |

检测各期货品种的数据集是否存在缺失值。对于含有大量缺失值的期货品种，本文将其剔除；若缺失值较少，则采取前向填充或后向填充的方式进行补全，以减少数据丢失对模型的影响。由于各个商品期货品种存在数量级差别，因此本文对其进行归一化处理。

## 3.2高相关性期货对获取

### 3.2.1主成分+OPTICS聚类

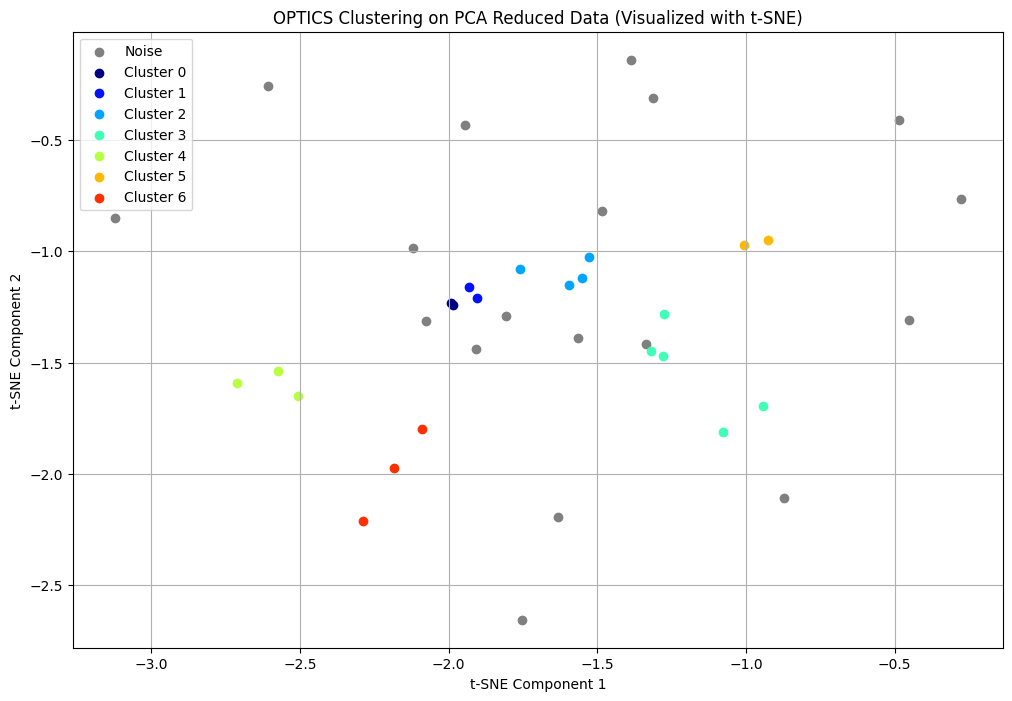
为了减少冗余信息并提取主要影响期货价格波动的因素，本文对期货的每日收益率进行了主成分分析。根据每日收盘价计算和上一收盘价之间的波动率，以每日波动率作为特征，进行主成分降维。



**图3.1 主成分分析方差解释图**

PCA的结果表明，前24个主成分解释了超过90%的数据方差，因此选择了这24个主成分用于后续分析。排序每个主成分对收益率的贡献，例如，第1主成分中2022年5月18日被识别为主要代表日期，该日期数据可以揭示回测时间段内期货市场的主要波动模式。

对降维后的数据进行OPTICS聚类。为了更直观地展示聚类结果，本文使用t-SNE将高维数据降维到二维，并对聚类结果进行了可视化。每个簇代表了一组波动特性相似的期货品种，而噪声点则以灰色标识，表明这些点未被分配到任何一个簇。

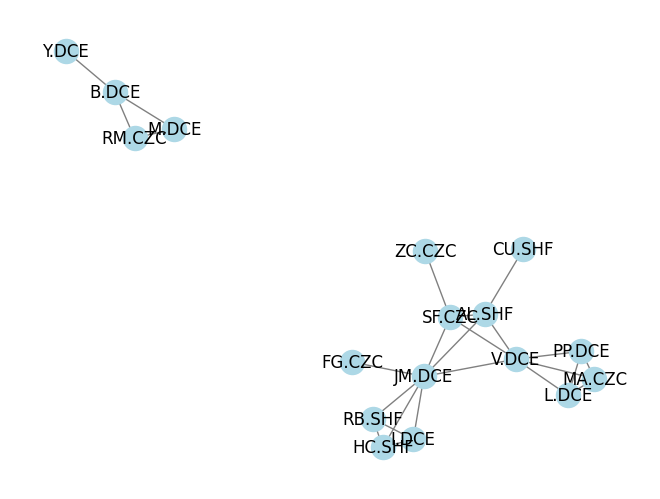


**图3.2 Optics聚类效果图**

Cluster2包含有色金属期货（如AL.SHF、CU.SHF、PB.SHF、ZN.SHF），表明这些金属期货在波动特性上高度相关；Cluster4包含农产品期货（如B.DCE、M.DCE、RM.CZC），显示了这些商品在波动性上的一致性；Cluster6则涵盖了油脂类期货（如OI.CZC、P.DCE、Y.DCE），表明这些期货品种之间的密切相关性，这些分类结果跟目前市场上主流的商品期货品种划相类似。

### 3.2.2网络模型搭建

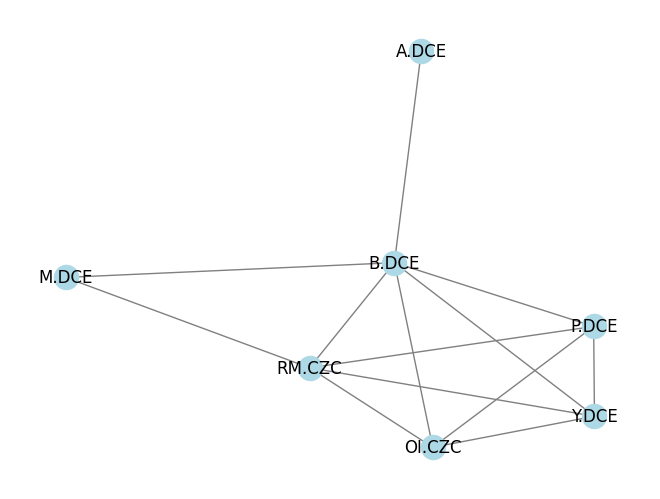
基于波动率计算相关性矩阵，并基于相关性阈值构建无向加权网络，模型根据阈值筛选出高波动率的期货，将波动率高度相关的期货品种通过边相连，结果如下：



**图3.3 全品种网络模型效果图**

图中展示了多个高度关联的集群。例如，HC.SHF、RB.SHF、IM.DCE、V.DCE之间形成了一个密集的子网络，这些品种之间的波动率具有很高的相关性。该集群可能代表某些市场板块或产业链内部的强关联性。AL.SHF、V.DCE和I.DCE作为中心节点连接了多个其他品种。与中心节点不同，ZN.SHF、CS.DCE这些节点位于网络的边缘，连接较少，波动性与其他期货品种的相关性相对较弱。

按照市场对商品期货的分类，单独选取农产品期货分析：



**图3.4 农产品期货网络模型效果图**

从农产品期货的波动率网络图中可以看到，多个农产品期货品种之间存在显著的波动性相关性。油脂类期货OI.CZC、Y.DCE和P.DCE形成了一个密集的三角形子网络。

基于聚类和网络综合分析，本文最终选择了以下几组期货品种进行配对交易：

（1）有色金属类：CU.SHF+AL.SHF、AL.SHF+ZN.SHF、CU.SHF+ZN.SHF；

（2）农产品类：B.DCE+M.DCE、B.DCE+RM.CZC、M.DCE+RM.CZC；

（3）油脂类：OI.CZC+P.DCE、OI.CZC+Y.DCE、P.DCE+Y.DCE。

这些期货品种对的选择是基于它们在网络中表现出的强相关性，特别是在农产品和有色金属期货之间的密切联系。通过选择这些品种进行配对交易，可以有效利用波动率之间的相关性进行套利，同时降低市场的风险暴露。这些配对品种不仅在波动性上具有较强的联动性，还可能为交易者提供更多的套利机会。

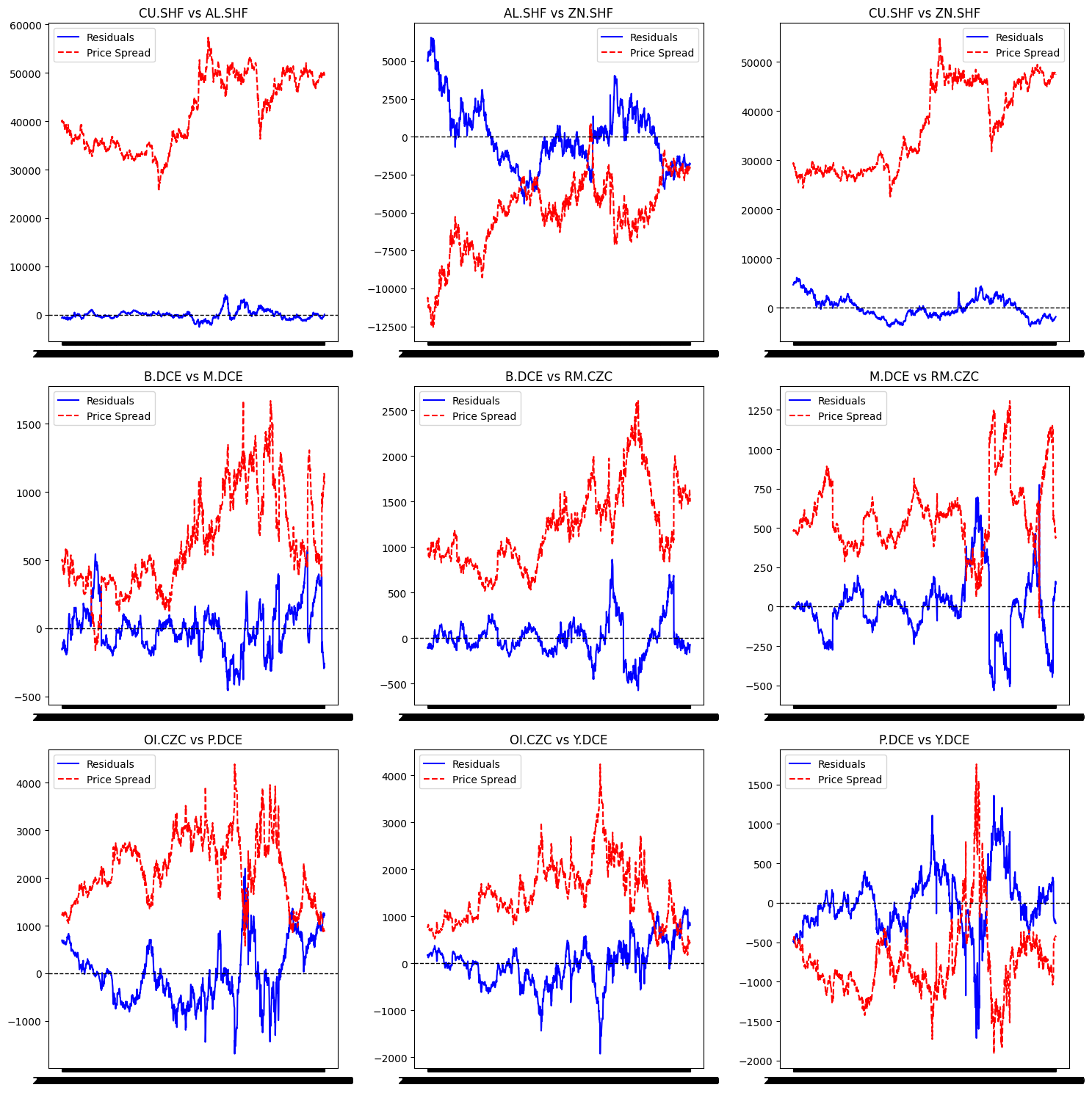
## 3.3协整性分析

观察各期货对的价格差和回归残差进行ADF单位根检验结果如下：

**表3.2各期货对的价格差和回归残差**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 期货对 | 价差平稳性 | 回归残差P值 | 回归残差平稳性 | 协整性 |
| CU.SHF+AL.SHF | 不平稳 | 0.0001 | 平稳 | 有 |
| AL.SHF+ZN.SHF | 不平稳 | 0.0590 | 不平稳 | 无 |
| CU.SHF+ZN.SHF | 不平稳 | 0.0837 | 不平稳 | 无 |
| B.DCE+M.DCE | 不平稳 | 0.0001 | 平稳 | 有 |
| B.DCE+RM.CZC | 不平稳 | 0.0024 | 平稳 | 有 |
| M.DCE+RM.CZC | 平稳 | 0.0182 | 平稳 | 无 |
| OI.CZC+P.DCE | 不平稳 | 0.0297 | 平稳 | 无 |
| OI.CZC+Y.DCE | 不平稳 | 0.0433 | 平稳 | 无 |
| P.DCE+Y.DCE | 平稳 | 0.0035 | 平稳 | 有 |

观察回归残差和价差的时间序列图如下：



**图3.5 全品种期货回归残差和价差的时间序列图**

基于协整性判断，结合CU.SHF和AL.SHF、B.DCE和M.DCE、B.DCE和RM.CZC、P.DCE和Y.DCE这些配对的残差表现出一定的均值回复特性，之后的策略研究将主要侧重这四对期货。

# 4 实证策略构建

## 4.1实证策略构建

根据3.3节协整性判断得到的回归方程，使用其估计的参数对期货对进行对冲价差模型配对交易。定义为配对交易第一腿的期货品种，为第二腿的期货品种，根据对冲比率α求出价差序列的均值：

将原价差序列减去均值，即去中心化得到序列，使去中心化后的序列均值为零。对去中心化后的价差序列求样本标准差，并制定交易规则如下（以CU.SHF和AL.SHF为例）：

（1）当时开仓，卖出1个单位头寸的CU.SHF主力合约，买入个单位头寸的AL.SHF主力合约，在时反向平仓；

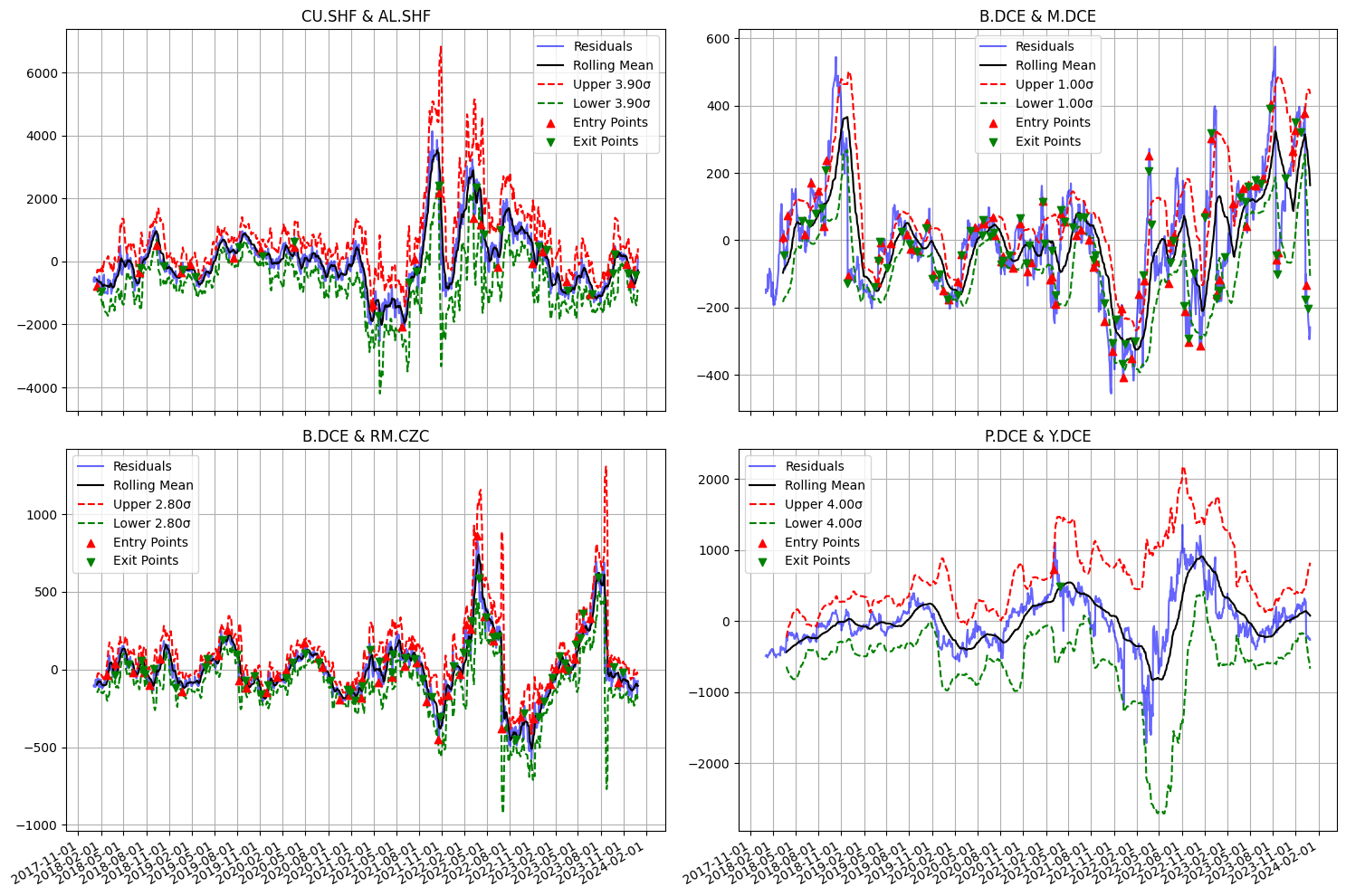
（2）当时开仓，买入1个单位头寸的CU.SHF主力合约，卖出个单位头寸的AL.SHF主力合约，在0时反向平仓。

## 4.2策略参数优化

在此处考虑时变的对冲比率，即通过交易当天向前倒退一个窗口期的数据进行对冲比率的计算，计算出的比率用作当天的交易策略的比率。将的值设定在（1，5）范围内每次步长为0.1；同时将标准差和对冲比率的窗口期范围设定为（10，60），每次步长为5。使用随机搜索的超参数优化，最终得到如下的实证结果：

**表4.1各期货对配对交易参数优化结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 配对交易对 | 窗口期 | n阈值 | 年化收益率 | 最大回撤 | 夏普比率 |
| CU.SHF&AL.SHF | 10 | 3.9 | 4.01% | 1.03% | 0.44 |
| B.DCE&M.DCE | 45 | 1.0 | 0.90% | 0.37% | -3.31 |
| B.DCE&RM.CZC | 10 | 2.8 | 0.50% | 0.65% | -5.76 |
| P.DCE&Y.DCE | 55 | 4.0 | 1.28% | 0.00% | -3.13 |



**图4.1 各期货对配对交易点位**

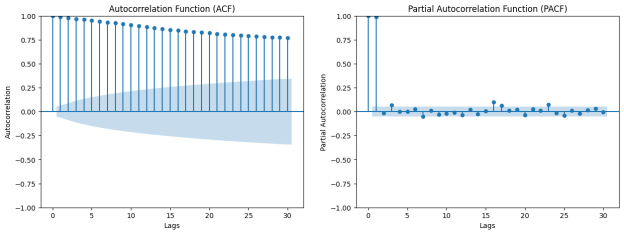
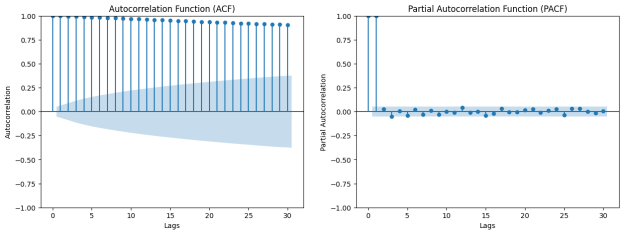
从四个配对交易策略的表现来看，CU.SHF&AL.SHF策略无疑是最优选择，展现出最高的年化收益率和较低的最大回撤，同时其夏普比率达到0.44，表明该策略在风险与收益之间实现了良好的平衡，适合用于长期套利交易。相比之下，其余虽然最大回撤较低，但由于收益不足以覆盖风险，表明策略稳定性不足且盈利能力欠佳。在市场波动较小或不确定性较高的情况下，这类策略可能更适合短期套利，但若要提升其实际价值，需要优化参数，如引入动态阈值调整和波动率过滤器，避免频繁交易导致收益恶化。

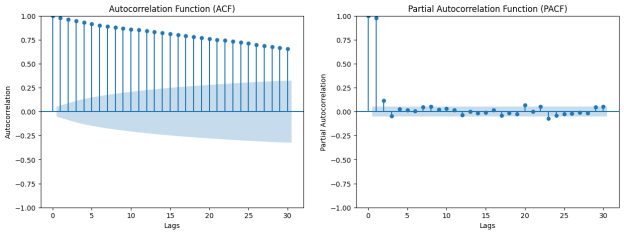
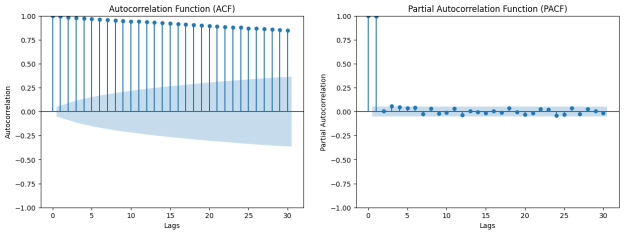
## 4.3时变标准差模型引入

### 4.3.1 ARIMA模型拟合

在一个ARIMA模型中，需要确定三个参数：p,q,d。ARIMA(p,d,q)可以看成是三个参数模型构建而来的组合模型：AR是自回归模型，p代表了AR模型的自回归参数；MA为滑动平均模型，q代表了MA模型的滑动平均参数，d为阶数，即作d次差分后序列平稳。

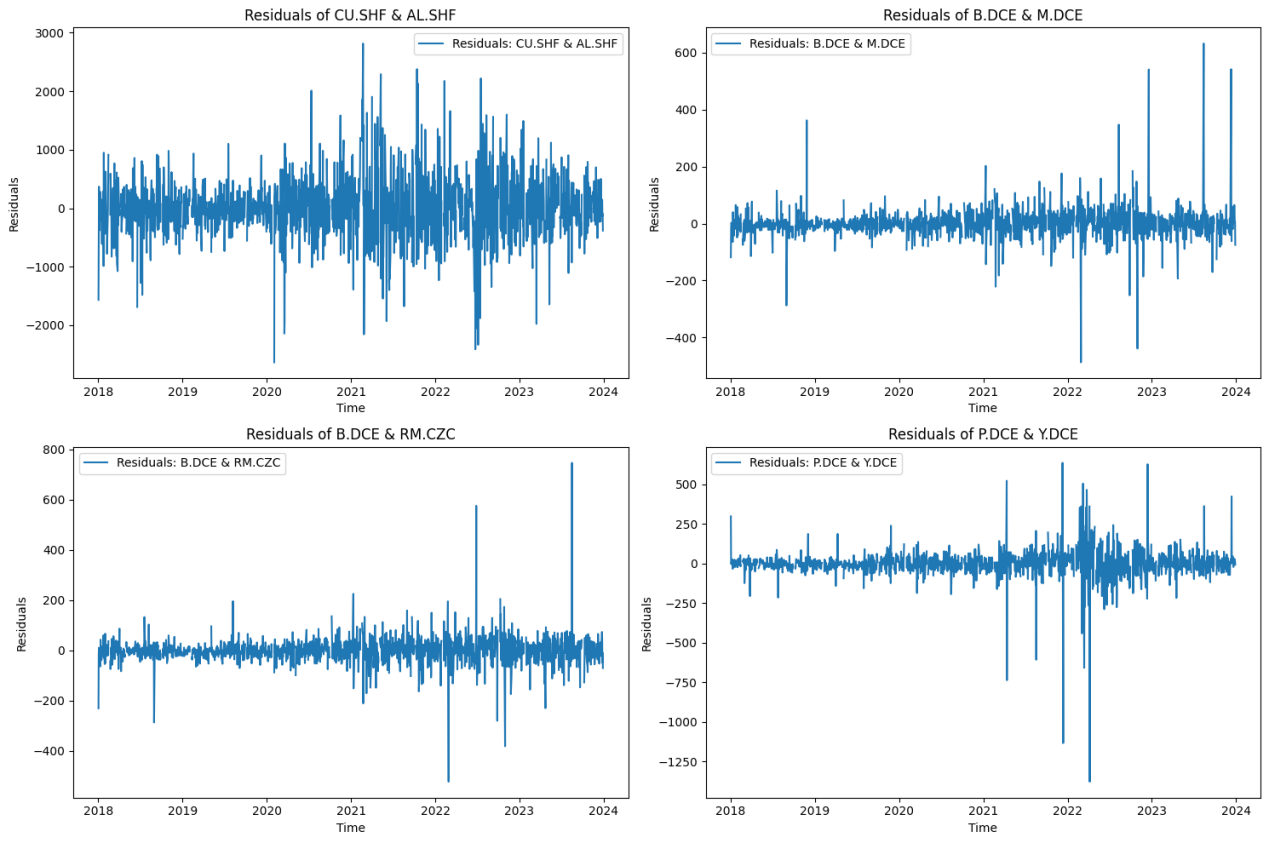
利用之前时变对冲策略计算出来的对冲比率，得到价差序列。在之前章节中已经证明价差序列是平稳的，因此的d即为0（拟合ARMA模型）。画出的价差序列的自相关、偏自相关图如下：





**图4.2 各期货对价差序列的自相关、偏自相关图**

一阶自相关呈现出拖尾现象，而偏相关图有类似于截尾的表达，因此拟合一个AR(p)都模型可能是一个更好的选择。采用AIC或BIC原则进一步去判断，即模型中AIC或BIC值越小，模型就越好。AIC/BIC原则均判断差价序列适合拟合AR(1)，观察残差的波动性，画出AR(1)模型残差序列图：



**图4.3 各期货对AR(1)模型残差序列图**

残差在中间部分波动性比较大，两端相比波动较小，存在为异方差的可能性，在之后的策略中构造GARCH(1,1)模型拟合异方差。

### 4.3.2策略制定及策略参数优化

利用求出的时变的异方差模型标准差以及去中心化后的价差序列，重复3.5.2，制定如下策略（以CU.SHF和AL.SHF为例）：

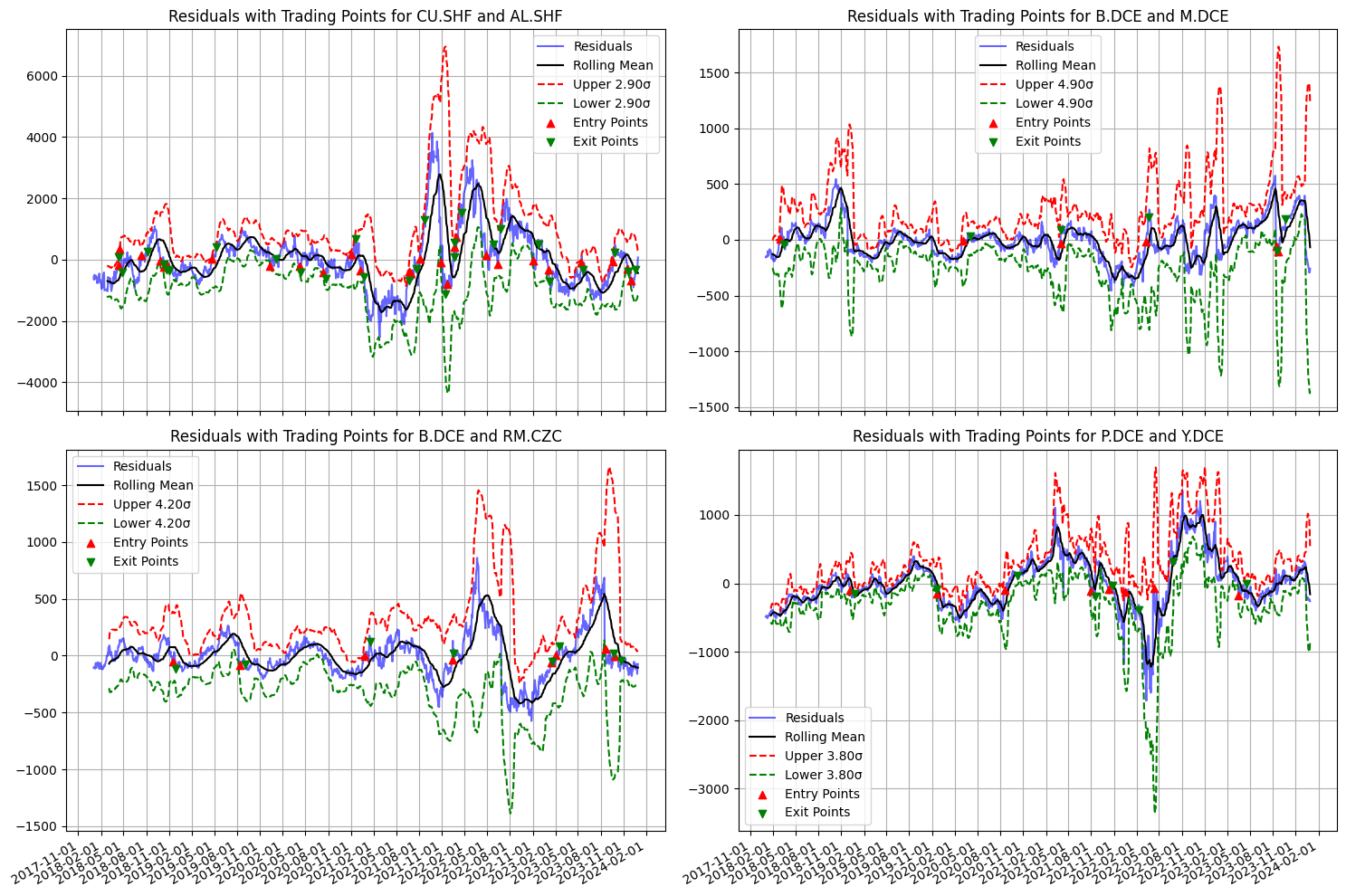
（1）当时开仓，卖出1个单位头寸的CU.SHF主力合约，买入个单位头寸的AL.SHF主力合约，在时反向平仓；

（2）当时开仓，买入1个单位头寸的CU.SHF主力合约，卖出个单位头寸的AL.SHF主力合约，在0时反向平仓。

重复4.2中的参数优化，将的值设定在（1，5）范围内每次步长为0.1；同时将标准差和对冲比率的窗口期范围设定为（10，60），每次步长为5。使用随机搜索的超参数优化，最终得到如下的实证结果：

**表4.2各期货对异方差配对交易参数优化结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 配对交易对 | 窗口期 | n阈值 | 年化收益率 | 最大回撤 | 夏普比率 |
| CU.SHF&AL.SHF | 10 | 3.0 | 2.93% | 1.21% | -0.05 |
| B.DCE&M.DCE | 45 | 1.0 | 0.48% | 0.51% | -4.76 |
| B.DCE&RM.CZC | 15 | 3.7 | 0.67% | 0.04% | -8.66 |
| P.DCE&Y.DCE | 10 | 1.1 | 0.10% | 0.92% | -4.41 |



**图4.4 各期货对异方差配对交易点位**

在上述四个配对策略中，CU.SHF&AL.SHF表现最佳，但伴随1.21%的回撤，显示出较高收益与一定风险并存的特点。B.DCE&RM.CZC尽管回撤控制良好，但年化收益率和夏普较低，表明该策略难以在当前市场环境中盈利。

## 4.4基于基差的策略配置

### 4.4.1基差研究

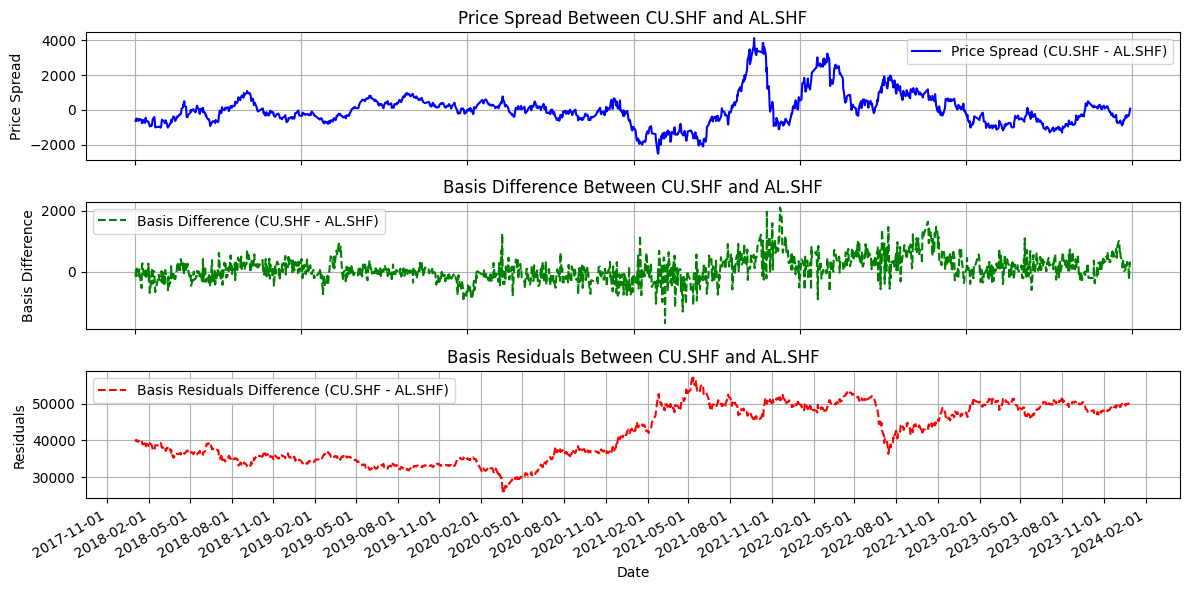
基差交易（Bias）是期货市场中广泛使用的一种策略，主要利用现货价格和期货价格之间的差异（即“基差”）进行操作。交易者通过预判基差的扩大或收缩来获利，这一策略在商品市场尤为常见，如农产品、贵金属以及金融衍生品。本文研究尝试使用基差代替价差进行配对交易，其原理基于如下推断。

定义为配对交易第一腿的期货品种，为第二腿的期货品种；表示为配对交易第一腿的现货品种，为第二腿的现货品种，前文中所使用的价差可以表示为：

本文认为该价差存在值回归特性。而此时新的价差表示方式为基差之差。即：

基于现货与期货价格的趋同，即随着期货合约到期，现货和期货的基差应逐渐收敛。因此，当基差出现明显的扩张或收缩时，本文可以预期它会回归到一个稳定水平，从而形成交易机会。基于基差之差的策略相比传统价差交易，更能反映市场供需、季节性波动和地域差异的影响。例如，在农产品市场中，不同地区的天气状况可能导致不同市场之间的基差差异，因此基差交易比单纯的价差交易更加灵活。

观测基差和之前的价差以及对冲后残差序列的相似性（以CU.SHF和AL.SHF为例）：



**图4.5 CU.SHF和AL.SHF价差、回归残差、基差序列图**

基差与价差存在一定的相似性，和回归残差却有较大的区别（其余品种均表现出相似的特性）。研究基差差的平稳性，发现所选择四个品种均平稳。

### 4.4.2策略配置及参数优化

利用4.4.1节中的代替4.1中的，制定如下策略（以P.DCE和Y.DCE为例）：

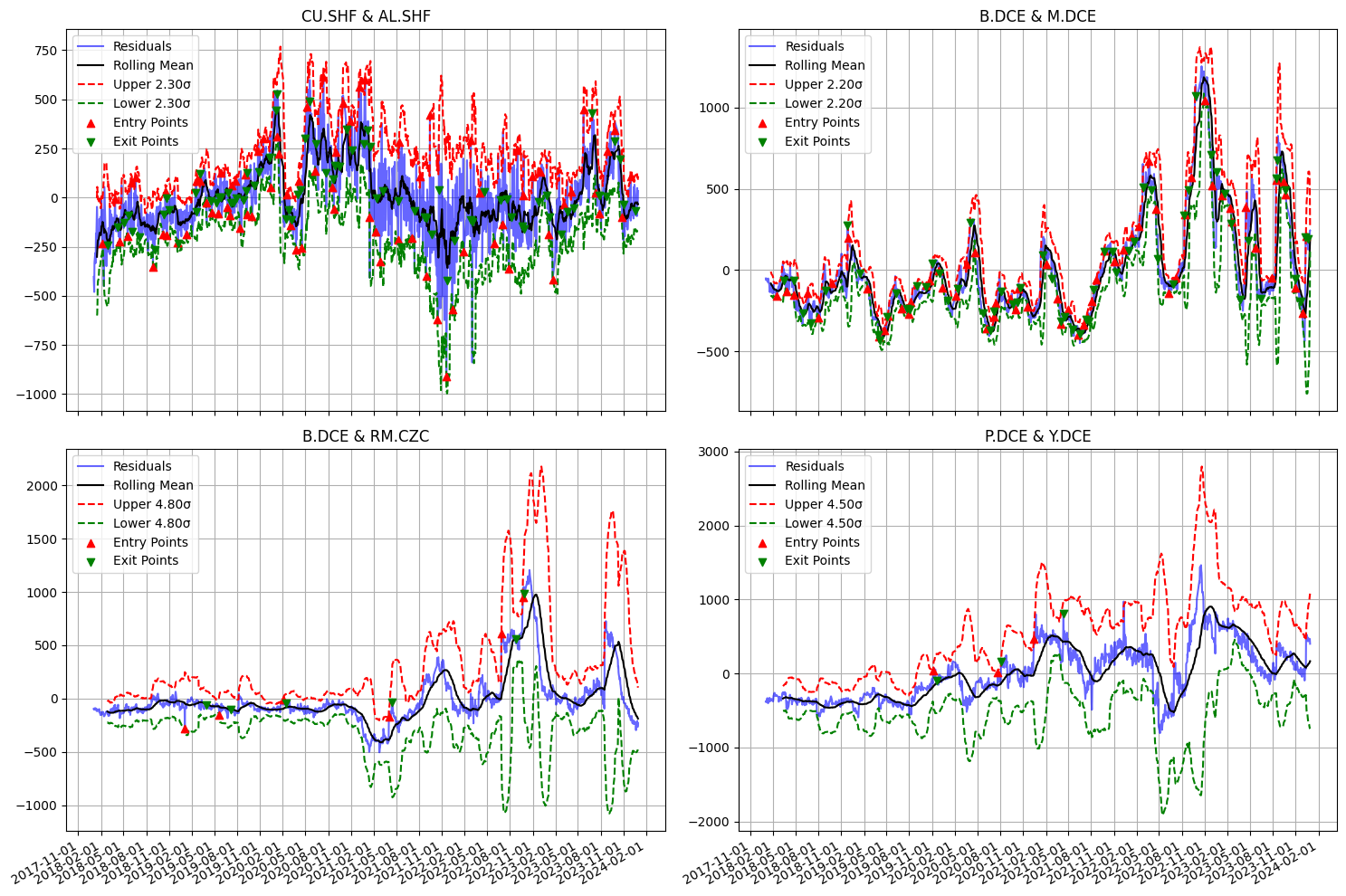
（1）当时开仓，卖出1个单位头寸的P.DCE主力合约，买入个单位头寸的Y.DCE主力合约，在时反向平仓；

（2）当时开仓，买入1个单位头寸的P.DCE主力合约，卖出个单位头寸的Y.DCE主力合约，在0时反向平仓。

重复4.2中的参数优化，将的值设定在（1，5）范围内每次步长为0.1；同时将标准差和对冲比率的窗口期范围设定为（10，60），每次步长为5。使用随机搜索的超参数优化，实证结果如下：

**表4.3各期货对基差配对交易参数优化结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 配对交易对 | 窗口期 | n阈值 | 年化收益率 | 最大回撤 | 夏普比率 |
| CU.SHF&AL.SHF | 10 | 2.3 | 6.47% | 6.25% | 0.49 |
| B.DCE&M.DCE | 15 | 2.2 | 1.25% | 0.52% | -2.12 |
| B.DCE&RM.CZC | 35 | 4.8 | 0.54% | 0.20% | -5.64 |
| P.DCE&Y.DCE | 45 | 4.5 | 1.65% | 0.00% | -1.93 |



**图4.6 各期货对基差配对交易点位**

从残差图和数据表现来看，配对交易策略在大部分时间内保持稳定，但在特定时期（如2021-2022年）出现明显波动，可能与市场不确定性事件相关。整体策略通过设定阈值（如2.3倍或4.8倍标准差）触发交易信号，并使用不同长度的窗口期平滑波动，确保噪声不会频繁触发交易。CU.SHF & AL.SHF配对表现最佳，年化收益率为6.47%，夏普比率为0.49，说明有良好的风险调整收益。其他配对，如B.DCE & M.DCE 和 P.DCE & Y.DCE，虽然回撤控制较好，但收益和夏普比率均为负，表明收益不足以弥补风险，策略效果不理想。此外，从残差波动来看，策略表现依赖于较长窗口期的观察，这也符合商品市场中季节性和周期性趋势的特性。

# 5 拓展策略构建

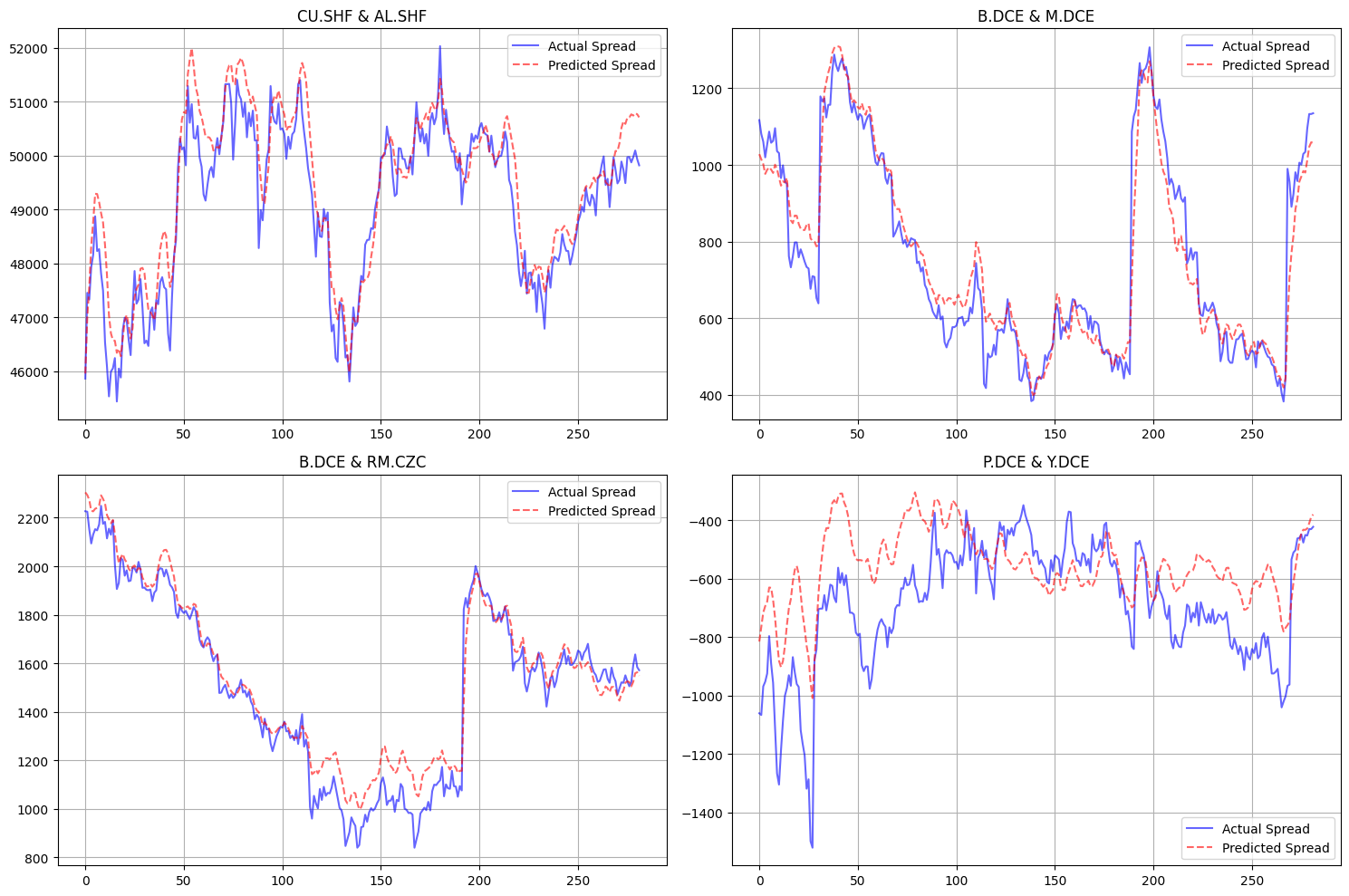
本节将使用机器学习的方法进行策略构建，但相较于第4节差距在于，本节将使用预测后的价格进行配对交易，即使用t时刻的特征预测t+1时刻的价格，判断t+1时刻的价差判断是否符合开仓条件，如果满足则在t时刻开仓。使用这样的策略相较于第4节中的策略，会更早一步判断市场的趋势，更快的捕捉到未来的价格偏离和均值回复，以及形成更高效的套利捕捉。

## 5.1机器学习效果

### 5.1.1 LSTM神经网络

通过LSTM神经网络对两个商品期货之间的价格差进行预测，包含以下特征输入：CLOSE，HIGH，LOW，VOLUME，AMT，PCT\_CHG，SWING，OI，ANAL\_BASIS，BASISANNUALYIELD，使用50天窗口期生成特征序列，使用MinMaxScaler将数据标准化到[0,1]范围，以确保所有特征的取值范围一致，防止模型训练过程中某些特征因量级不同而主导模型。

第一层LSTM设计为100个单元，保留中间层的时间序列信息；第二层LSTM为100个单元，不返回序列，仅输出最后一步的结果，输出层为单个神经元。损失函数使用均方误差，优化器为Adam。训练参数epochs=20，使用小批量梯度下降，每次更新参数时处理32个样本。模型参数经过超参数优化，训练参数确保模型loss不再大幅下降，已达到了最佳的训练效果，预测结果如下：

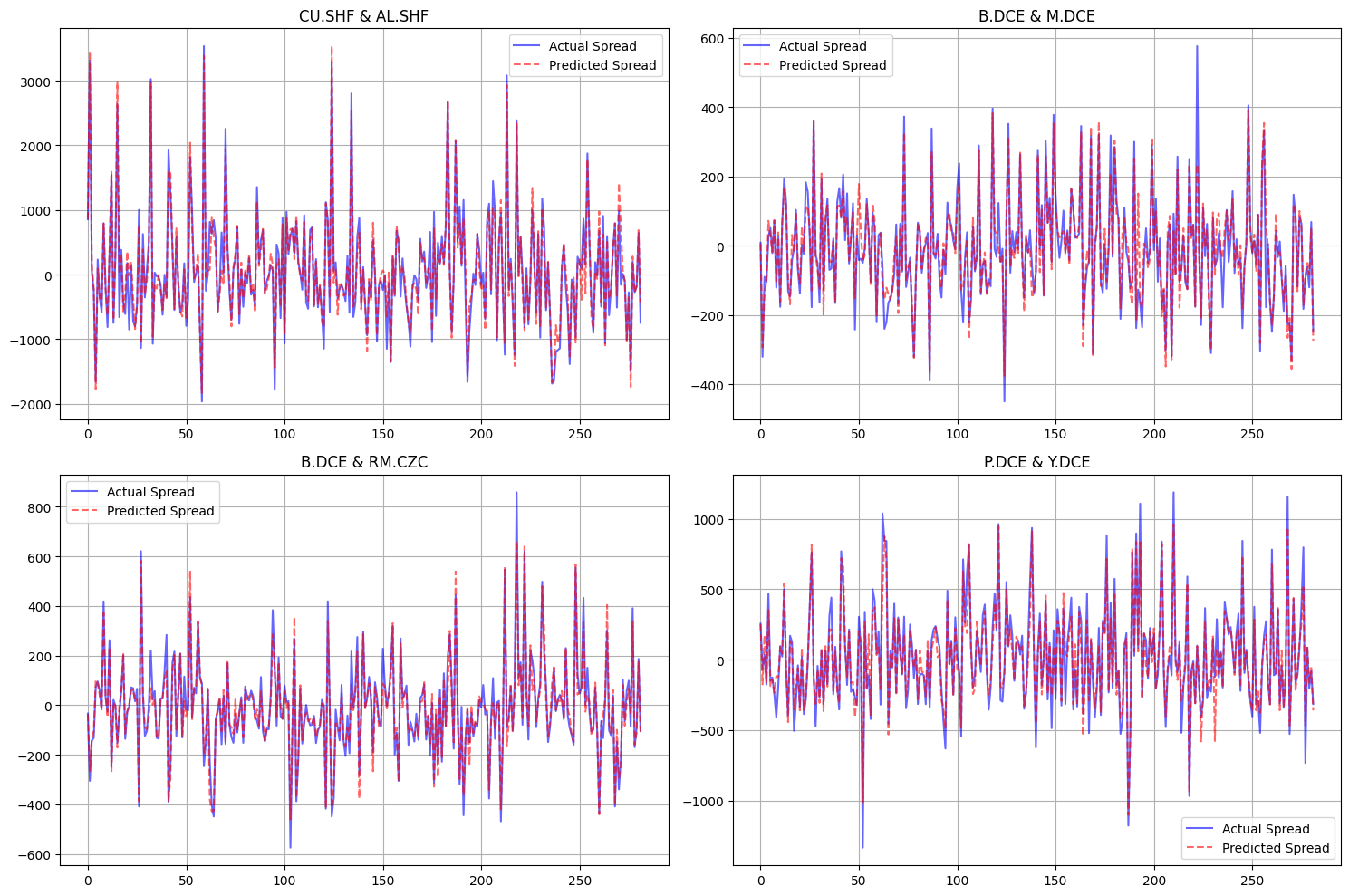


**图5.1 LSTM神经网络预测效果图**

从图中可以看出，LSTM模型在预测CU.SHF 和 AL.SHF期货价差方面表现出了一定的趋势捕捉能力，但在Y.DCE和P.DCE的样本中预测较不精确。

### 5.1.2 XGboost

XGBoost回归模型同样选取相同的数据特征，调参优化后的参数具体为使用100棵树来提升模型的精度，学习率设置为0.2，平衡模型收敛速度与精度，树的最大深度为6，以防止模型过拟合，同时保证其捕捉到复杂的非线性关系。模型采用了均方误差作为损失函数，并对训练数据进行拟合，预测结果如下：



**图5.2 XGboost预测效果图**

XGBoost模型能够较好地跟踪市场的短期波动，但在处理剧烈变动和长期趋势方面表现欠佳，相较于LSTM更适合作为一种短期波动预测工具，因此对于根据价格波动而导致的套利机会可能会有更好的表现。

## 5.2机器学习效果实证

使用机器学习预测的价格计算出预测后的，根据模仿4.1中的策略构造方法（以P.DCE和Y.DCE为例）：

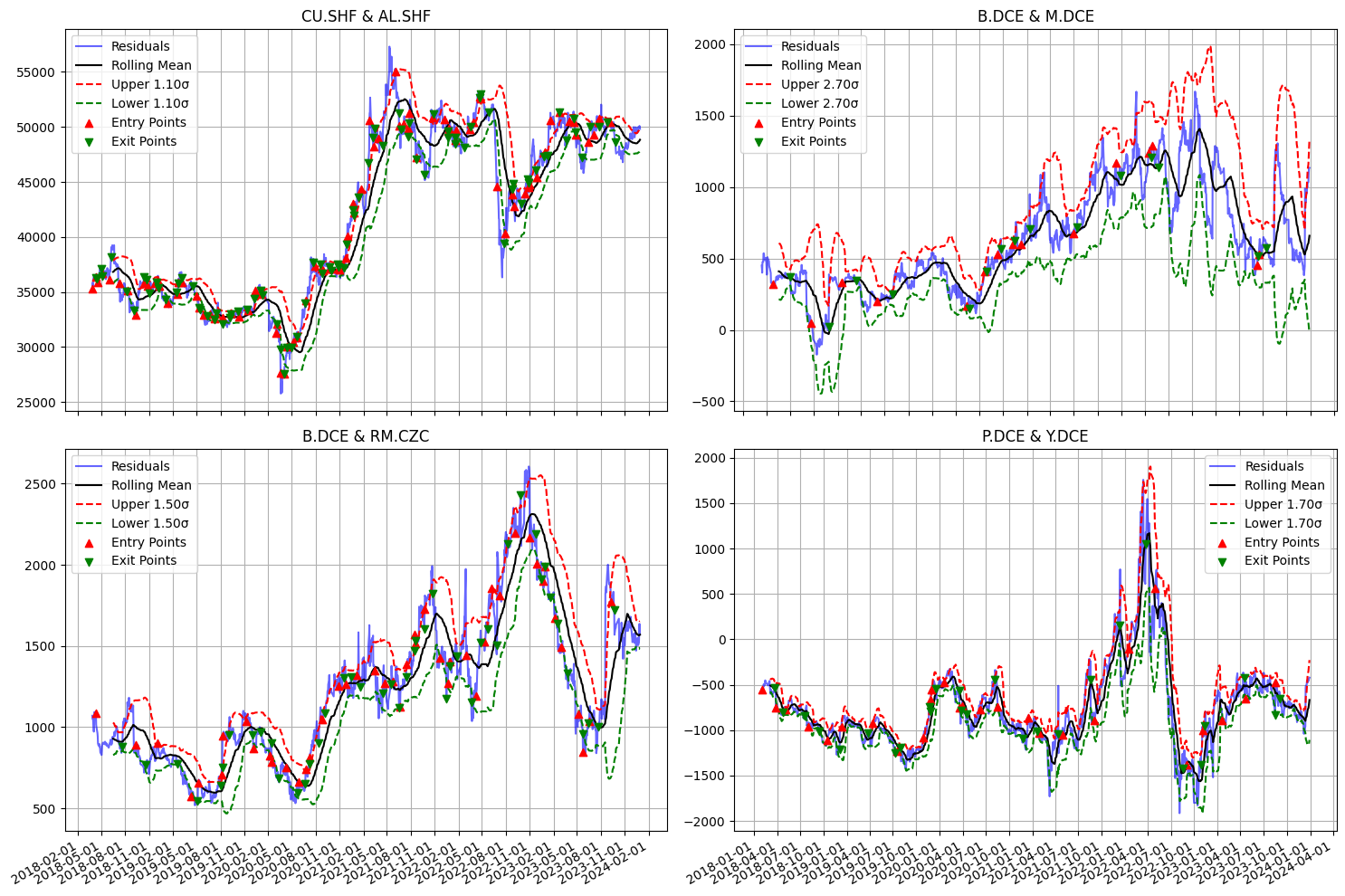
（1）当时开仓，在t时刻卖出1个单位头寸的P.DCE主力合约，买入个单位头寸的Y.DCE主力合约，在时反向平仓；

（2）当时开仓，在t时刻买入1个单位头寸的P.DCE主力合约，卖出个单位头寸的Y.DCE主力合约，在0时反向平仓。

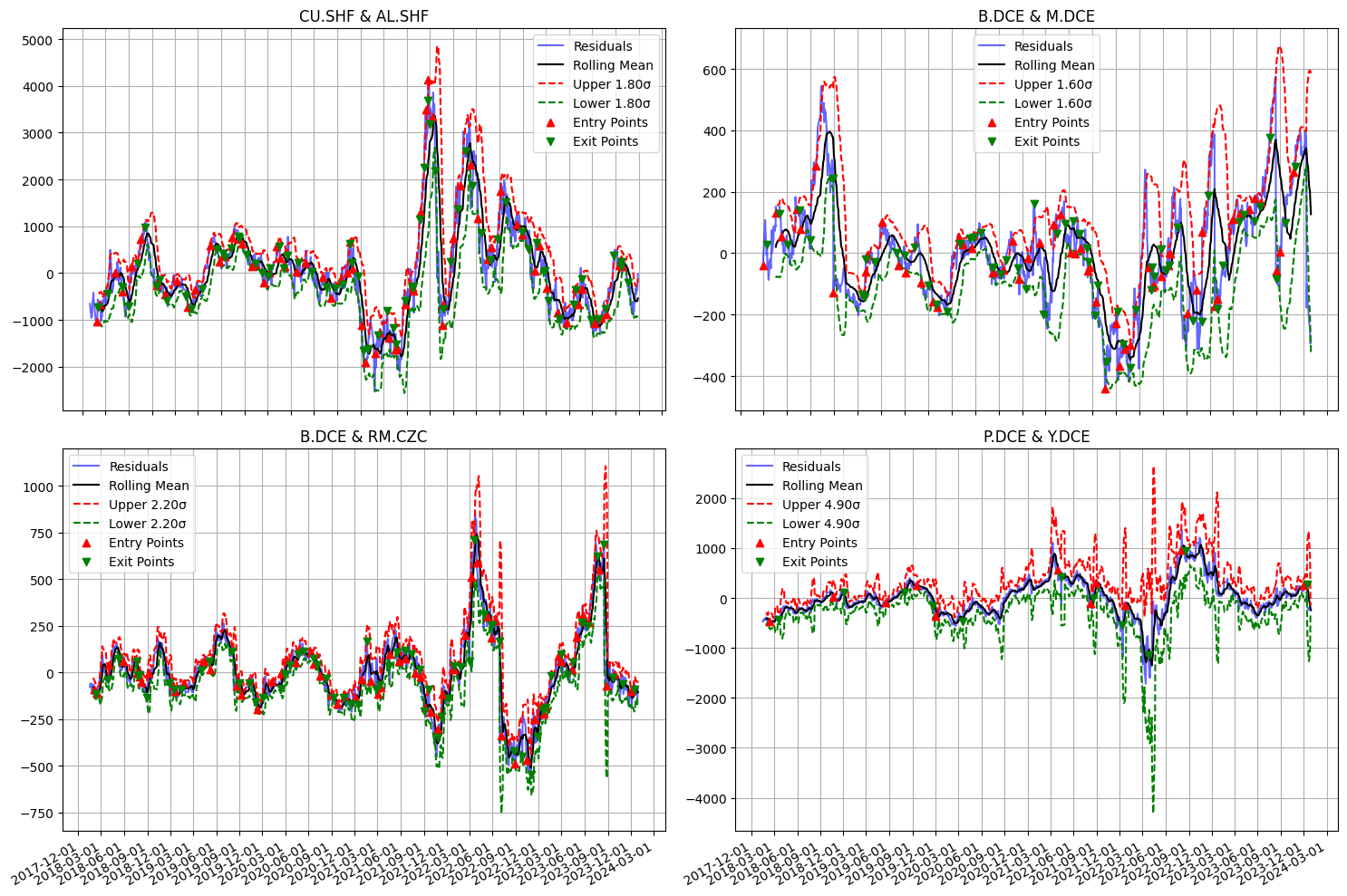
同样的使用随机搜索的方法，对窗口期和阈值进行优化，得到最终的优化结果为：

**表5.1 机器学习下的配对交易参数优化结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 配对交易对 | 模型 | 年化收益率 | 最大回撤 | 夏普比率 |
| B.DCE & M.DCE | LSTM | 0.57% | 0.25% | -6.22 |
| XGBoost | 0.65% | 1.25% | -2.45 |
| B.DCE & RM.CZC | LSTM | 1.11% | 0.45% | -2.64 |
| XGBoost | 0.34% | 0.37% | -5.4 |
| CU.SHF & AL.SHF | LSTM | 3.75% | 0.97% | 0.22 |
| XGBoost | 5.77% | 0.00% | 0.91 |
| P.DCE & Y.DCE | LSTM | 3.22% | 0.10% | 0.11 |
| XGBoost | 1.65% | 0.00% | -1.93 |



**图5.1 LSTM神经网络配对交易位点图**



**图5.2 XGBoost配对交易位点图**

XGBoost和LSTM在CU.SHF和AL.SHF配对交易的回测结果中展现了各自的优势。XGBoost 模型实现了 年化收益率 5.77% 和 夏普比率 0.91，并且最大回撤为 0.00%，表现出极高的短期交易稳定性。这表明 XGBoost 在捕捉市场中的短期波动和实现低风险套利方面具备一定优势，适合应用于频繁交易的策略。

相比之下，LSTM 模型取得了 年化收益率 3.75% 和 夏普比率 0.22。虽然整体收益不及 XGBoost，但其在 45 天窗口期内保持了稳定表现，更适合捕捉较长周期的趋势性机会。此外，LSTM 在最大回撤 0.97% 的水平上有效控制了风险，展示了在处理长期趋势中的稳健性。总体来看，XGBoost 更适合于捕捉短期市场波动、实现高频套利，而 LSTM 在较长周期的趋势跟踪上展现了稳健的特性。两者可结合使用，在策略中同时融入长期和短期信号，从而进一步提高交易策略的稳定性与综合收益表现。

# 6 总结与展望

## 6.1 总结

本文以配对交易策略为研究核心，系统地探讨了不同金融模型和数据处理技术在期货市场中的应用，通过对2018年1月1日至2023年12月31日的多个大宗商品期货数据进行分析，验证了其在中国市场中的有效性。研究采用了协整分析、主成分分析（PCA）、OPTICS聚类、条件异方差模型（GARCH），并结合了LSTM和XGBoost等机器学习技术，通过PCA和OPTICS算法，对多个商品期货的波动率数据进行降维和聚类，成功筛选出高度相关的期货组合。

基于协整检验，本文发现部分商品对，如CU.SHF与AL.SHF、P.DCE与Y.DCE，具备长期均衡关系，并适合作为配对交易标的。采用去中心化价差构建策略，通过开仓和平仓时点的标准差阈值判断进行交易。利用GARCH模型捕捉金融时间序列中的波动性集聚现象，为交易阈值提供动态调整依据。基于LSTM和XGBoost对未来价差进行预测，帮助捕捉市场趋势和短期波动。测试结果表明，XGBoost模型在短期套利上表现最佳，而LSTM则在捕捉长期趋势上更具优势。

对于CU.SHF和AL.SHF，本文所构建的几种交易策略结果如下：

**表6.1 各策略在CU.SHF和AL.SHF品种上的回测效果比较**

|  |  |  |  |
| --- | --- | --- | --- |
| 策略 | 年化收益率 | 最大回撤 | 夏普比率 |
| 对冲价差模型 | 4.01% | 1.03% | 0.44 |
| GARCH模型 | 2.93% | 1.21% | -0.05 |
| 基差模型 | 6.47% | 6.25% | 0.49 |
| LSTM模型 | 3.75% | 0.97% | 0.22 |
| XGBoost模型 | 5.77% | 0.00% | 0.91 |

对冲价差模型实现了年化收益率4.01%、最大回撤1.03%和夏普比率0.44，展现出较好的收益与风险平衡，适合用于相对稳健的套利策略。引入异方差后的GARCH模型的年化收益率2.93%和夏普比率-0.05，表明其在市场波动性捕捉上有所优势，1.21%的最大回撤反映了较高的风险，使其更适用于波动性较大的市场环境，有可能是ARIMA模型拟合导致。基差模型在各策略中表现突出，实现了6.47%的年化收益率，但由于6.25%的较高回撤，其收益波动较大。尽管如此，该模型通过基差动态反映市场的错位，提供了良好的套利机会。LSTM模型展示了出色的长期稳定性，年化收益率为3.75%、最大回撤仅为0.97%、夏普比率0.22，适合捕捉长期趋势信号，但在短期收益上不如其他策略。XGBoost模型取得了5.77%的年化收益率和0.91的夏普比率，且实现了最小最大回撤，在各模型中风险控制能力最强，特别适合用于高频交易和短期套利策略。

## 6.2 不足与展望

尽管本文展示了配对交易策略在中国期货市场中的潜力，但仍存在一些局限性和未来可以改进的方向。

(1) 交易成本的考虑不足

本文的回测结果未包含交易成本（如佣金、滑点、税费等），这些费用在真实交易中会显著影响策略的收益表现。

未来应引入真实市场交易成本模型，并优化交易频率，以避免因过度交易导致的收益下降。同时，结合市场流动性分析，确保策略在高频环境中的可执行性。

(2) 特征变量的扩展

本文仅使用了收盘价、波动率等基础数据进行建模，未充分利用开盘价、交易量、持仓量等其他关键特征。

后续研究可以通过构建更多交易指标，如资金流动指标、成交量均衡指标等，进一步提升模型的预测能力。

(3) 机器学习模型的进一步优化

当前研究中使用的LSTM和XGBoost模型对特征工程和超参数优化的依赖较大。未来可以引入Transformer网络等更为先进的深度学习模型，以提高对复杂市场数据的理解和趋势捕捉能力。同时，探索混合模型（如LSTM和XGBoost的组合）以兼顾短期和长期交易机会，优化策略表现。

# 参考文献

1. Gatev E, Goetzmann W N, Rouwenhorst K G. Pairs Trading: Performance of a Relative-Value Arbitrage Rule[J]. The Review of Financial Studies, 2006, 19(3): 797-827.
2. 钱小安. 对套利和图利的研究[J]. 国际金融研究, 1995, (01): 36-39.
3. 毛二万, 宋逢明. 风险、收益、套利和估价[J]. 系统工程学报, 2000, (04): 333-337.
4. Melamed L. The mechanics of a commodity futures exchange: A critique of automation of the transaction process[C]. Hofstra Law Review, 1977.
5. Kanamura T, Rachev S T, Fabozzi F J. A profit model for spread trading with an application to energy futures[I]. The Journal of Trading, 2010, 5(1): 48-62.
6. Simon D P. The soybean crush spread: Empirical evidence and trading strategies[J]. Journal of Futures Markets, 1999, 19(3): 271-289.
7. Ross S A. The arbitrage theory of capital asset pricing[J]. Journal of Economic Theory, 1976.
8. Gatev E, Goetzmann W N, Rouwenhorst K G. Pairs Trading: Performance of a Relative-Value Arbitrage Rule[J]. The Review of Financial Studies, 2006, 19(3): 797-827.
9. Qazi L T, Rahman A U, Gul S. Which Pairs of Stocks should we Trade? Selection of Pairs for Statistical Arbitrage and Pairs Trading in Karachi Stock Exchange[J]. The Pakistan Development Review, 2015, 54(3): 215-244.
10. Goncu A, Akyıldırım E. Statistical Arbitrage with Pairs Trading[J]. International Review of Finance, 2016, 16(2): 307-319.
11. Avellaneda M, Lee J. Statistical arbitrage in the US equities market[J]. Quantitative Finance, 2010, 10: 761-782.
12. Krauss C, Do X A, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017, 259: 689-702.
13. Lam J W, You Y. Machine learning statistical gravity from multi-region entanglement entropy[J]. Physical Review Research, 2021.
14. Senn S. Francis Galton and Regression to the Mean[J]. Significance, 2011, 8(3): 124-126.
15. Zakamulin V. Secular mean reversion and long-run predictability of the stock market[J]. Bulletin of Economic Research, 2016. DOI:10.1111/boer.12105.
16. 贺洪增. 基于复杂网络的股票市场稳定措施和板块预测研究[D]. 北京科技大学, 2023.
17. 乔舰, 范淑芬. 时间序列分析中几个概念的解析[J]. 高等数学研究, 2023, 26(01): 57-61.
18. Hamilton J D. Time Series Analysis[M]. Princeton University Press, 1994.
19. Engle R F, Granger C W J. Co-integration and Error Correction: Representation, Estimation, and Testing[J]. Econometrica, 1987, 55(2): 251-276.
20. 赵耀. 基于协整理论与阈值动态化的银行股配对交易策略的优化研究[J]. 技术与市场, 2022, 29(09): 164-166+169.
21. Engle R F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation[J]. Econometrica, 1982, 50(4): 987-1007.
22. 张静瑄. 基于Copula模型的股指期货统计套利策略设计[D]. 西北师范大学, 2022.
23. 李静. 基于强化学习算法的商品期货配对交易策略设计[D]. 上海师范大学, 2021.
24. 刘辉, 刘忠元, 周伟杰. 基于随机过程与支持向量机构建期货配对交易策略[J]. 常州大学学报(社会科学版), 2018, 19(03): 59-64.
25. 杜文静. 基于协整的豆粕期货跨期套利组合研究[J]. 商展经济, 2021, (03): 57-60.
26. 由林青.配对交易策略中交易信号动态化研究[D].北京工商大学,2018.