

商品期货统计套利策略研究

(申请清华大学应用统计硕士专业学位论文)

培养单位： 数学科学系

申请人： 张文瀚

指导老师： 唐宏岩副教授

二〇一九年五月

Research on Statistical Arbitrage Strategy of Commodity Futures

Thesis submit to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Applied Statistics

by

Zhang Wenhan

Thesis Supervisor: Associate Professor Tang Hongyan

May, 2019

摘要

商品期货市场有着双向交易、高杠杆的特点，其风险要比股票市场更大，但收益也更为可观。面对期货市场高风险的特点，许多期货投资者希望找到一种策略，能够在取得超额收益的同时，对冲掉大部分的市场风险，而统计套利策略就能够较好地满足投资者的这一需求。配对交易是统计套利策略中的一种，在配对交易中，投资者同时做多和做空两种价格高度相关的资产，利用两种资产价差的均值回复特性来获取收益。

本文使用 2015-2017 年的商品期货价格数据进行研究，从商品期货组合的流动性、现实逻辑等几个角度综合考虑，选取了四组商品期货套利组合。经过协整检验，我们确定了这几个组合的价差序列具有平稳性，并通过对其价差序列进行交易实现了统计套利策略。本文使用了样本外的 2018 年数据对策略进行回测，结果发现大多数组合都有着夏普比较高，回撤较小的优点。将几个策略按照等资金的方式组合后，发现策略的最大回撤和年化波动率进一步下降。与单个套利策略相比，组合策略的风险更小，夏普比更高。

关键词：协整模型；统计套利；商品期货

Abstract

Commodity futures market has the characteristics of two-way trading and high leverage. Its risk is greater than that of stock market, but its return is also more considerable. Faced with the high risk of the futures market, many futures investors hope to find an investment strategy that can hedge most of the market risk while obtaining excess return. Statistical arbitrage strategy can meet the needs of these investors. Pairs trading is one of the statistical arbitrage strategies. In pairs trading, investors long one asset and short another assets that is highly related to it. Investors use the mean recovery characteristics of the spread between two assets to obtain return.

In this thesis, We use commodity futures price data from 2015 to 2017 to research the investment strategy. Four pairs of commodity futures arbitrage portfolios are selected from the perspective of liquidity and realistic logic of commodity futures portfolio. Then, this thesis test these futures price data by cointegration and confirm the stationary of spread time series. In this thesis, we use the data of 2018 outside the sample to measure the strategy. The results show that most portfolios have the advantages of higher Sharp ratio and lower withdrawal. After combining several strategies with equal funds, it is found that the maximum withdrawal and annual volatility of the strategies are further reduced. Compared with single arbitrage strategy, portfolio strategy has less risk and higher Sharp ratio.

Keyword: Cointegration model; Statistical Arbitrage; commodity futures

目录

第 1 章 引言	1
1.1 研究背景	1
1.1.1 统计套利策略简介	1
1.1.2 商品期货简介	2
1.2 文献综述	3
1.3 论文主要工作与结构安排	5
第 2 章 相关理论	6
2.1 $AR(p)$ 时间序列模型	6
2.2 时间序列模型的单位根检验	6
2.3 协整理论	7
第 3 章 商品期货统计套利策略	9
3.1 数据选取与处理	9
3.2 策略建立与回测的基本流程	9
3.3 套利对象的选取	10
3.3.1 商品期货的流动性	10
3.3.2 期货价格相关性	12
3.3.3 平稳性检验与协整检验	12
3.4 期货组合配比的确定	14
3.5 开仓与平仓阈值的确定	15
第 4 章 策略回测结果分析	17
4.1 交易成本参数设定	17
4.2 策略评价指标	17
4.3 策略回测结果分析	18
第 5 章 结论	21
参考文献	22
致谢	23
声 明	24
个人简历、在学期间发表的学术论文与研究成果	25

第1章 引言

1.1 研究背景

1.1.1 统计套利策略简介

量化投资，是指投资者通过建立数学模型将投资理念数量化，并借助计算机程序进行自动化交易来获取收益的投资方式。与传统的投资方式相比，量化投资有着对历史数据的应用更加充分，交易更加迅速，主观因素影响更小的优点。量化投资兴起于上世纪80年代，至今已有30余年的历史。而我国的量化投资行业则起步较晚，目前仍处于高速发展的阶段。

统计套利是一类重要的量化投资策略，其主要思路是：通过对历史数据的分析得出一定的统计规律，并根据规律构建投资组合，发现真实市场对于资产组合的错误定价，并从中获取收益。由于统计套利是基于过去的统计规律，统计套利策略具有一定的风险性，但通过承担一定的风险能够获得可观的收益。

Hogan et al. (2004)^[1] 对统计套利策略进行了数学定义：如果一个零初始成本的自融资交易策略 $(x(t): t \geq 0)$ 的累计收益折现值为 $v(t)$ ，且 $v(t)$ 满足：

$$(1) \quad v(0) = 0, \quad (1-1)$$

$$(2) \quad \lim_{t \rightarrow +\infty} E[v(t)] > 0, \quad (1-2)$$

$$(3) \quad \lim_{t \rightarrow +\infty} P[v(t) < 0] = 0, \quad (1-3)$$

$$(4) \quad \text{若对于任意 } t < \infty, P[v(t) < 0] > 0, \text{ 则 } \lim_{t \rightarrow +\infty} \text{Var}[v(t)]/t = 0.$$

则称这是一个统计套利策略。其中第一条要求初始收益为0；第二条要求当经过的时间充分长时，策略的收益期望大于0；第三条要求当时间趋于无穷时，策略收益为负的概率收敛于0；第四条要求如果在任意有限时间内策略亏损的概率均大于0，那么当时间趋于无穷时，策略收益方差与时间之比的极限为0，这可以理解为统计套利策略可能出现亏损，但随着时间的积累，策略的风险应当可控。

一般来说，统计套利策略包括配对交易、多因子套利、波动率交易等若干不同的类型，而其中最经典的一种则是配对交易。20世纪80年代，摩根士丹利的研究人员提出了“黑匣子”策略，通过寻找价格相互关联的一组股票来进行配对交易，

并取得了丰厚的收益。时至今日，配对交易策略依然被广泛应用于量化投资当中，是一种行之有效的策略。

匹配交易的核心逻辑是价格的均值回复规律：如果两种资产具有较高的关联，并且其价格具有相似的历史行为，那么这两种资产的价差就可能具有均值回复的特性，价差在偏离均值一段时间后有较大概率会回复到均值水平。基于这一规律，该策略在价差偏离均值一定程度时进行反向操作以获取收益。具体地，匹配交易策略的实现包含三个步骤：一是选择两种价格波动相互关联的资产A和B，作为交易标的；二是确定一个开仓阈值，并在两种资产的价差超过这一阈值时，做多价格偏低的资产并做空价格偏高的资产；三是在两种资产的价差回归均值时，平仓并获得收益。匹配交易策略同时做多和做空一组高度相关的资产，从而对冲掉了大部分的市场风险，使得策略的风险相对可控。

匹配交易要求所选择的一组交易标的即可以被做多，也可以被做空。而在中国市场，做空股票一般来说比较困难，因此通常选择期货作为匹配交易的标的资产。本文主要研究商品期货市场中的匹配交易策略，通过构建若干组商品期货价差组合来实现统计套利。

1.1.2 商品期货简介

期货合约，是指双方约定在未来的某一时刻以合约约定的价格买入或卖出某一产品的合约。而商品期货合约，则是指双方交易的产品为实物商品的期货合约。大约在13世纪，欧洲地区已经出现了双方约定在未来进行商品交割的交易形式，这可以看作是期货合约的雏形。1848年，美国芝加哥地区的商人建立了芝加哥期货交易所，并逐步发展出最早的标准化期货合约。而中国的期货市场发端于20世纪90年代，经国务院批准，郑州粮食批发市场于1990年引入期货交易机制，标志着国内对期货市场探索的开始。目前中国主要的商品期货交易所包括上海期货交易所，大连期货交易所，郑州期货交易所这三大交易所，上市交易的期货产品包括金属类、农副产品类、化工类、能源类这几大类。

与股票交易相比，期货交易的两个重要特点是买卖双向交易和保证金制度。买卖双向交易是指交易者可以自由选择作为期货合约的买方或卖方。因此期货市场即可以做多也可以做空，无论期货价格上涨还是下跌交易者都有机会获得收益。而保证金制度则要求交易者按照合约价格的一定比例缴纳保证金，并在出现亏损后继续追加保证金或选择平仓。通常国内商品期货的保证金比例为5%-10%，因此期货交易有着较高的杠杆，交易者可以用较少的资金完成大宗的交易。在期货交易中，大

部分交易者的目的并非在未来买入或卖出标的商品，而是从期货价格的涨跌中获取收益。因此大部分期货交易者并不会等待合约到期后进行交割，而是会选择在合约到期前进行平仓，以获取投资收益。

一个标准化的期货合约一般包含期货的交割时间，保证金比例等重要信息。下表是一份 2017 年的螺纹钢期货合约所包含的主要信息。

表 1.1 螺纹钢期货合约主要信息

交易品种	螺纹钢
交易单位	10 吨/手
报价单位	元（人民币）/吨
最小变动价位	1 元/吨
涨跌停板幅度	上一交易日结算价±3%
合约交割月份	1—12 月
交易时间	上午 9:00—11:30，下午 1:30—3:00 和 交易所规定的其他交易时间
最后交易日	合约月份的 15 日
交割日期	最后交易日后连续五个工作日
最低交易保证金	合约价值的 5%
交割方式	实物交割
交易代码	RB
上市交易所	上海期货交易所

1.2 文献综述

研究统计套利策略的学术文献数量众多，其中主要的建模方法包括距离法，协整法和时间序列法这几大类，下面我们对这几类方法进行简单的介绍。

研究统计套利策略模型的第一类方法是距离法，以 Gatev et al. (2006)^[2] 为代表。这篇文章选择了 1962–2002 年的美国股市的全部流动股票进行研究，并用 P_{it} 表示持有第 i 支股票到时间 t 的累计总收益。文中将股票 i 与股票 j 之间的距离定义为：

$$SSD_{ijt} = \frac{1}{T} \sum_{t=1}^T (P_{it} - P_{jt})^2. \quad (1-4)$$

作者将样本中全部的 n 支股票两两配对形成 C_n^2 个股票对，并对每个股票对计算距离 SSD_{ijt} ，选取其中距离最小的若干个股票对进行交易。交易规则为在价差偏离值达到两倍标准差时开仓进行交易，在价差回归均值时平仓。这样的方法能够取得一定的收益，例如距离最小的5个股票对的年化超额收益为5.98%。但这样的配对方式不够精确，Do&Faff(2010)^[3]指出，将 Gatev et al. (2006)方法的使用期延长至2009年，策略的收益率就会有大幅衰减，这是因为其中有32%的股票价差不会收敛于历史均值。Do&Faff(2010)认为，许多配对的价差不收敛是由于所选择的股票本身关联并不密切，因此策略缺乏充足的现实逻辑。只对同一个行业内的股票进行配对可以提高价差收敛于均值的股票对比例，进而提高策略收益。

总体来说，距离法是一种简单且比较有效的配对交易方法，其主要的问题是按照距离来挑选股票对，会忽视掉价差大但价差均值稳定的股票对；距离小意味着价差波动率比较小，这样挑选会错过价差均值稳定但波动率高的高收益组合。

研究统计套利策略模型的第二类方法是协整模型法，Vidyamurthy(2004)^[4]是使用这类方法的文献中较为早期的一篇。Vidyamurthy(2004)的做法基于APT模型，将股票的收益分解为公共因子收益和特异因子收益。而挑选配对交易对象的原则，就是寻找具有相似公共因子暴露的股票。Vidyamurthy(2004)的方法并没有用到严格的协整检验，但其中的主要思想就是协整方法，按照这种方法挑选出的股票对也具有良好的协整关系。

Vidyamurthy(2004)的方法相对比较复杂，而 Dunis et al.(2006)^[5]采用的是直接对两个价格序列进行协整检验的方法，也取得了不错的效果。他对WTI原油期货和布伦特原油期货的价格进行了协整检验，并根据回归系数确定了期货组合配比，并在样本外数据取得了28%的收益和1.63的夏普比。

与距离法相比，协整模型法用到了更多的统计学思想，并对期货价差序列进行了时间序列建模，这也使得协整模型法更容易与其他时间序列方法相结合。通过协整模型法确定的价差组合具有平稳性，这意味着价差的均值和标准差应当趋于稳定，价差在偏离均值一段时间后较大的概率回复均值，因此更容易产生套利机会。本文采用的方法主要基于协整模型法。

研究统计套利策略模型的第三类方法是时间序列法，Elliott et al.(2005)^[6]是其中的一个代表。这篇文章假设两种资产 P 和 Q 的价差 $X_t = P_t - Q_t$ 具有均值回复的特性，并服从下面的Ornstein-Uhlenbeck随机过程：

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t. \quad (1-5)$$

其中选用 O-U 过程是为了刻画价差的均值回复特性。在上述模型假设下，文中提出的交易策略为：当 $X_t \geq \mu + c$ 时，做空资产 P 并做多资产 Q ；当 $X_t \leq \mu - c$ 时，做多资产 P 并做空资产 Q 。其中 c 为待确定的交易阈值，根据最大化收益率或最大化夏普比的交易目标分别进行求解。

与前两种相比，时间序列方法更好地刻画了统计套利策略所需的均值反转特性。但模型的设定也更为复杂，有可能出现过拟合的问题。

1.3 论文主要工作与结构安排

本文主要研究基于协整模型法的统计套利策略在中国商品期货市场中的应用。本文第一章主要介绍统计套利策略和国内商品期货市场的一些基本情况，并对统计套利策略的相关文献进行了综述。第二章介绍本文需要用到的基本理论。第三章使用 2015-2017 年的商品期货价格数据对统计套利策略进行设计，具体包括选择期货组合，确定期货组合配比，确定开仓和平仓阈值等几个步骤。第四章在样本期内（2015-2017 年）和样本期外（2018 年）对策略进行了回测，并对回测结果进行分析。第五章对回测结果进行总结，并提出可能的改进方案。

第2章 相关理论

2.1 AR(p)时间序列模型

自回归时间序列模型一般简称为AR模型，是一种常用的拟合平稳时间序列的模型。设我们要拟合的时间序列为 $\{x_t\}$ ，而 $\{\varepsilon_t\}$ 为一组零均值白噪声序列，则我们称具有以下结构的时间序列模型为 p 阶自回归模型，简记为AR(p)模型：

$$\begin{cases} x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} + \varepsilon_t, \\ \beta_p \neq 0, \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0 \quad s \neq t, \\ E(x_s \varepsilon_t) = 0 \quad \forall s < t. \end{cases} \quad (2-1)$$

AR(p)模型有三个附加条件。其中第一个条件要求 $\beta_p \neq 0$ ，保证了模型的阶数为 p ；第二个条件要求序列 $\{\varepsilon_t\}$ 为一个白噪声序列；而第三个条件要求新一期的随机扰动 ε_t 与之前各期的 x_s 无关。

2.2 时间序列模型的单位根检验

对于式(3-1)中的AR(p)时间序列模型，我们称如下多项式为模型的特征方程：

$$\lambda^p - \beta_1 \lambda^{p-1} - \beta_2 \lambda^{p-2} - \cdots - \beta_p = 0. \quad (2-2)$$

模型的特征方程的 p 个复根 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 被称为AR(p)模型的 p 个特征根。当且仅当模型的 p 个特征根均在单位圆内时，AR(p)模型具有平稳性。因此为了检验一个自回归模型是否具有平稳性，只需检验模型是否有单位根，这种检验方法被称为单位根检验法，也称为ADF检验法。

如果模型存在单位根 $\lambda_i = 1$ ，则由(3-3)式可得：

$$\beta_1 + \beta_2 + \cdots + \beta_p = 1. \quad (2-3)$$

因此为了检验模型的平稳性，我们可以定义 $\rho = \beta_1 + \beta_2 + \cdots + \beta_p - 1$ ，并检验 $\rho = 0$ 是否成立，那么我们可以将单位根检验的统计假设确定为：

$H_0: \rho = 0$ （等价于序列 $\{x_t\}$ 非平稳）

$H_1: \rho < 0$ （等价于序列 $\{x_t\}$ 平稳）

而ADF检验统计量为

$$\tau = \frac{\hat{\rho}}{S(\hat{\rho})}, \quad (2-4)$$

其中 $S(\hat{\rho})$ 为参数 ρ 的样本标准差。

在ADF检验中，如果检验结果显著拒绝了原假设，则说明时间序列 $\{x_t\}$ 显著平稳，此时也称序列 $\{x_t\}$ 为零阶单整序列。而如果检验结果未能拒绝原假设，则说明 $\{x_t\}$ 不具有平稳性，此时我们可以对序列 $\{x_t\}$ 进行差分运算，并检验差分序列的平稳性。如果序列 $\{x_t\}$ 至少需要经过 d 次差分运算才能得到一个平稳序列，则称序列 $\{x_t\}$ 是 d 阶单整序列。

在本文中，我们将检验所选取的套利组合中两种期货的价差是否具有平稳性，并对具有平稳性的价差组合使用 $AR(p)$ 模型进行拟合。

2.3 协整理论

传统的计量经济学研究当中，对两组时间序列建立线性回归模型的前提是两组时间序列均具有平稳性。如果自变量时间序列和因变量时间序列不具有平稳性，那么就可能会出现“虚假回归”问题，即回归方程并不能反映自变量和因变量间的逻辑关系，只是一种巧合。因此在对两组时间序列建立回归模型之前，应当先对它们的平稳性进行检验。

在一些实际问题当中，自变量序列和因变量序列均不满足平稳性，但序列之间却确实存在着一定的长期均衡关系，在这些问题当中回归模型也有可能是有意义的。为了刻画非平稳序列间的长期均衡关系，Engle 与 Grange 于 1987 年提出了协整理论。

如果我们在非平稳的自变量时间序列 $\{x_{1t}\}, \{x_{2t}\}, \cdots, \{x_{kt}\}$ 与因变量时间序列 $\{y_t\}$ 间构造了回归模型：

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i x_{it} + \varepsilon_t. \quad (2-5)$$

并且满足回归模型的残差 ε_t 具有平稳性，则称因变量序列 $\{y_t\}$ 与自变量序列 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 之间具有协整关系。在自变量序列与因变量序列不具有平稳性的条件下，如果自变量和因变量间具有协整关系，回归模型就仍然是有意义的，能够反映自变量与因变量间的内在联系。

为了避免虚假回归问题，在对多元时间序列建立线性回归模型时，必须先对自变量序列和因变量序列进行平稳性检验，如果自变量与因变量不满足平稳性再进行协整检验，在自变量与因变量具有平稳性或存在协整关系时才可以建立回归模型。目前较为常用的协整检验方法是 Engle&Grange (1987)^[7]提出的 Engle-Granger 两步法。（以下简称 E-G 两步法）

E-G 两步法的统计假设为：

H_0 ：时间序列 $\{y_t\}$ 与 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 间不存在协整关系（即残差序列 $\{\varepsilon_t\}$ 非平稳）

H_1 ：时间序列 $\{y_t\}$ 与 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 间存在协整关系（即残差序列 $\{\varepsilon_t\}$ 平稳）

E-G 两步法分两步对上述假设进行检验：

步骤一：建立自变量序列 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 与因变量序列 $\{y_t\}$ 间的线性回归模型：

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i x_{it} + \varepsilon_t. \quad (2-6)$$

步骤二：对步骤一中的回归残差序列 $\{\varepsilon_t\}$ 进行平稳性检验。在这一步中常用的平稳性检验法是单位根检验，本文也将通过这种方法进行平稳性检验。

第3章 商品期货统计套利策略

3.1 数据选取与处理

本文选取了 2015 年至 2017 年上海期货交易所和大连期货交易所共 28 种商品期货的每日价格数据来建立统计套利策略并进行回测，数据来源为锐思数据库。选取的商品期货品种如下表所示。

表 3.1 研究涉及的商品期货品种

交易所名称	选取商品期货品种
上海期货交易所	铜、铝、锌、铅、黄金、白银、螺纹钢、 线材、热轧卷板、燃料油、石油沥青、天 然橡胶
大连期货交易所	黄大豆一号、黄大豆二号、豆粕、豆油、 鸡蛋、玉米、玉米淀粉、棕榈油、焦煤、 焦炭、胶合板、聚丙烯、聚氯乙烯、聚乙 烯、铁矿石、纤维板

在处理期货价格数据的过程中，我们需要考虑期货交易的流动性和时限性问题。在每个期货品种的所有合约中，一般只有几个月份的合约交易量较大，而其他月份的合约则流动性较差。以螺纹钢为例，每年的螺纹钢期货中交割时间为 1 月、5 月、和 10 月的合约的成交量最大，流动性最好，我们称之为主力合约。在交易策略中，我们应当选取各个时段的主力合约进行交易。由于期货合约具有时效性，我们需要在每个合约到期前换手到下一个主力合约，以实现连续持仓。与这种主力合约换手的交易方式相对应，我们将期货每年各时期的主力合约价格数据拼接在一起，形成一个主力连续合约价格序列。在之后的研究中，我们全部采用期货的主力连续合约价格作为研究数据。

3.2 策略建立与回测的基本流程

配对交易策略的主要思想是寻找若干对价格具有相似历史行为的资产组合，并对每个资产组合的价差进行交易。具体地，配对交易策略的建模与回测主要包含一

下几个步骤。

1. 确定套利对象

我们首先根据商品期货组合在现实生活中的关联和组合的价格相关性确定若干对期货组合，然后对每个组合中两种期货的价格序列进行协整检验，两种期货的价格满足协整关系意味着其价差具有平稳性，比较适合进行套利交易。

2. 确定两种资产的配比

在通过协整检验后，我们对每组两种期货的价格序列进行线性回归，根据线性回归的系数确定配对交易中两种资产的配比。

3. 确定策略的开仓、平仓阈值

策略的开仓、平仓阈值是指当两种期货的价差达到何值时进行开仓、平仓操作。配对交易策略的开仓和平仓阈值一般根据价差的均值和标准差计算得出。

4. 利用样本外数据（2018 年的期货价格数据）对策略进行回测。

本文中的统计套利策略是基于 2015–2017 年的期货价格数据来建立的，为了确保回测的有效性，我们利用样本外的 2018 年数据对策略进行回测，并根据年化收益率、夏普比等若干指标对策略进行评价。

3.3 套利对象的选取

3.3.1 商品期货的流动性

在确定套利对象的过程中，我们需要考虑一下四个方面：一是选取的套利对象应当具有良好的流动性；二是每个套利组合中的两个品种应当在生产和生活中存在一定的关联，即我们的策略应当有一定的现实逻辑；三是每个套利组合中两种期货产品的价格相关性应当较高，四是每个套利组合中两种期货产品的价格序列应当满足协整关系。我们首先根据流动性和现实逻辑挑选出四个期货组合，然后对这些组合的价格相关性和协整关系进行检验。

如果我们选取的期货流动性较差，就会导致在实际交易中出现较大的滑点，甚至较大金额的交易无法完成，严重影响策略的有效性。因此，我们在选取套利对象的过程中，应当选取具有良好流动性的交易标的，排除掉成交量和成交额很少的期货品种。表 3-X 计算了 2015 年上海期货交易所和大连期货交易所各期货品种的日平均成交额（单位为万元）。从表中我们可以看到，排在最后八名的线材、黄大豆二号、燃料油、纤维板、胶合板、聚氯乙烯、热轧卷板和铅的日平均成交额均小于 10 亿，流动性要明显差于其他品种，因此我们在选择套利对象时首先排除这八个品种。

表 3.2 2015 年各期货品种日平均成交额（单位：万元）

排名	期货名称	日平均成交额	排名	期货名称	日平均成交额
1	铜	8297793	15	铝	515166
2	螺纹钢	8111118	16	黄大豆一号	481928
3	天然橡胶	6894641	17	焦煤	437723
4	铁矿石	6766124	18	玉米	433906
5	白银	5946822	19	鸡蛋	405369
6	豆粕	4899318	20	玉米淀粉	348751
7	黄金	4614777	21	铅	43821.5
8	聚乙烯	3730302	22	热轧卷板	31082.1
9	棕榈油	3639218	23	聚氯乙烯	27992.9
10	豆油	3512425	24	胶合板	6591.56
11	聚丙烯	2654315	25	纤维板	1542.14
12	锌	1764619	26	燃料油	382.545
13	焦炭	945917	27	黄大豆二号	89.8786
14	石油沥青	551639	28	线材	2.49218

按照商品在生产生活中的联系，流动性较好的二十种期货产品大致可以分为四类：

1. 黑色系类：螺纹钢、铁矿石、焦炭、焦煤

这四种产品均与钢铁行业和煤炭行业相关，而煤炭行业又是钢铁行业的上游产业，煤炭是炼钢的重要原材料，因此这四种产品联系较为紧密。

2. 其他金属类：铜、白银、黄金、锌、铝

其中铜、铝、锌在生产和生活中的用途有一定交集。而黄金、白银属于贵金属，与另外三种金属关联较小，但黄金和白银关联较大。

3. 农副产品类：豆粕、棕榈油、豆油、黄大豆一号、玉米、鸡蛋、玉米淀粉

其中豆粕、豆油和黄大豆一号都属于黄豆及豆制品，价格必然高度相关。玉米和玉米淀粉也是原料和制成品的关系，其价格关联较大。

4. 能源化工类：天然橡胶、聚乙烯、聚丙烯、石油沥青

这四种产品在加工合成与实际应用上都存在一定的关联。其中聚乙烯和聚丙烯的化学性质和功能都有类似之处，应当是最合适的套利组合。

由于同一类种的不同商品普遍存在关联，我们如果在一类中选出多个套利组合，

就会导致不同组合间的独立性较差，不利于策略的整理收益。因此我们从这四类中各选出两个联系较为紧密的期货产品进行配对交易，并在下一小节中对每个组合的价格序列进行相关性和协整检验。

表 3.3 商品期货套利组合选取

期货组合编号	期货品种一	期货品种二
1	螺纹钢	焦炭
2	铝	锌
3	玉米	玉米淀粉
4	聚乙烯	聚丙烯

3.3.2 期货价格相关性

我们首先求出每个期货组合中两种产品价格序列的相关系数，结果如下表所示。我们所选取的四个组合的价格相关系数均在 0.9 以上，说明每个组合的价格均高度相关，适合进行套利交易。

表 3.4 四组商品期货组合的价格相关性

期货组合编号	期货品种一	期货品种二	相关系数
1	螺纹钢	焦炭	0.9429
2	铝	锌	0.9020
3	玉米	玉米淀粉	0.9528
4	聚乙烯	聚丙烯	0.9231

3.3.3 平稳性检验与协整检验

协整检验是统计套利策略研究中的常用方法。期货价格时间序列一般不具有平稳性，我们不容易用时间序列的建模方法对其建模。但如果两个期货产品的价格具有协整关系，那么对这两种期货的价格进行线性回归得到的残差就具有平稳性。我们可以将两个价格序列线性回归的残差看作套利交易的交易标的，并对其进行建模。

我们首先使用单位根检验法对选取的期货组合中的8个期货品种进行平稳性检验，检验结果如下表所示。八种商品的检验结果均未能拒绝原假设，说明每种商品的价格序列均不平稳。

表 3.5 期货价格平稳性检验

期货品种编号	期货品种	t-Statistic	p-value
1	螺纹钢	-0.5786	0.8757
2	焦炭	-0.9899	0.7569
3	铝	-0.8891	0.7916
4	锌	0.0759	0.9644
5	玉米	-1.7054	0.4283
6	玉米淀粉	-1.4856	0.5406
7	聚乙烯	-2.7327	0.0686
8	聚丙烯	-1.6120	0.4769

接下来，我们对这八种产品价格的一阶差分序列进行平稳性检验，发现这八种期货价格的一阶差分序列都具有平稳性，因此这八种期货的价格时间序列都是一阶单整的。这八种期货是同阶单整的，说明每组组合中可能存在协整关系。

表 3.6 期货价格一阶差分平稳性检验

期货品种编号	期货品种	t-Statistic	p-value
1	螺纹钢	-27.0641	0.0000
2	焦炭	-7.7148	1.2361e-11
3	铝	-14.9918	1.1223e-27
4	锌	-17.3942	4.9889e-30
5	玉米	-25.1318	0.0000
6	玉米淀粉	-25.8912	0.0000
7	聚乙烯	-17.3816	5.0687e-30
8	聚丙烯	-19.5354	0.0000

最后，我们对每个组合中的两个产品进行协整检验，检验结果见表 3.7。在 0.2 的显著性水平下，四个组合都拒绝了原假设，通过了协整检验。因此我们可以认为四个组合都具有协整关系，每组两个期货产品的线性回归残差是平稳序列。我们将

两个产品的回归残差看作我们要进行套利的价差，构造资产组合进行套利。

表 3.7 四组商品期货组合的协整检验结果

期货组合编号	期货品种一	期货品种二	协整检验 p 值
1	螺纹钢	焦炭	0.1105
2	铝	锌	0.1306
3	玉米	玉米淀粉	0.0653
4	聚乙烯	聚丙烯	0.0648

3.4 期货组合配比的确定

经过协整检验，我们发现每个期货组合中的两个产品都具有协整关系，因此它们的线性回归残差具有平稳性。下面我们对每个组合中的两个期货产品价格进行线性回归，并根据线性回归的系数构造多空资产组合，使得资产组合的价格等于线性回归的残差。这样构造的资产组合价格具有平稳性，我们可以对其进行套利交易。

设第 $i(1 \leq i \leq 4)$ 个期货组合中的第一个品种为 y_i ，第二个品种为 x_i 。 t 时刻第 i 个期货组合中第一个品种的价格为 y_{it} ，第二个品种的价格为 x_{it} 。我们对时间序列 $\{y_{it}\}$ 和 $\{x_{it}\}$ 建立线性回归模型：

$$y_{it} = \beta_{i1}x_{it} + \beta_{i0} + \varepsilon_{it}. \quad (3-1)$$

由时间序列 $\{y_{it}\}$ 和 $\{x_{it}\}$ 具有协整性，可得线性回归的残差序列 $\{\varepsilon_{it}\}$ 具有平稳性。我们构造投资组合 $z_i = y_i - \beta_{i1}x_i$ ，即做多 1 单位品种 y_i ，做空 β_{i1} 单位品种 x_i 。则投资组合 z_i 在 t 时刻的价格为 $\beta_{i0} + \varepsilon_{it}$ 。由 ε_{it} 的平稳性，投资组合 z_i 的价格会围绕均值 β_{i0} 上下波动，这适合我们进行套利交易。

下表是对期货品种 y_i 和 x_i 的价格序列进行线性回归后得到的结果。我们根据回归系数 β_{i1} 的取值来确定两个期货品种的配比，并构建投资组合。

表 3.8 期货价格序列回归结果

序号 <i>i</i>	y_i	x_i	β_{i0}	β_{i1}	R^2
1	螺纹钢	焦炭	1054.1940	1.2596	0.8892
2	铝	锌	6908.2100	0.3308	0.8136
3	玉米	玉米淀粉	246.1440	0.8586	0.9078
4	聚乙烯	聚丙烯	3913.8391	0.6574	0.8522

3.5 开仓与平仓阈值的确定

关于开仓与平仓的阈值问题，Vidyamurthy(2004)一书中通过蒙特卡洛模拟的方式得出了一条经验规则：假设价差时间序列减去均值是一个白噪声序列，那么在价差偏离均值 0.75 倍标准差时开仓，在价差偏离均值 2 倍标准差时进行止损，能够使得收益最大化。本文将参照上述的经验结果来确定策略的开仓与平仓阈值。表 3-X 对确定开仓与平仓阈值所需的价差投资组合均值和标准差进行了计算。其中 $z_i(1 \leq i \leq 4)$ 是我们所选定的 4 组价差投资组合， μ_i 表示价差投资组合的均值， σ_i 表示价差投资组合的标准差。

表 3.9 价差投资组合的均值、标准差计算

序号 <i>i</i>	价差投资组合 z_i	投资组合均值 μ_i	投资组合标准差 σ_i
1	$z_1 = y_1 - 1.2596x_1$	1.2596	228.0916
2	$z_2 = y_2 - 0.3308x_2$	0.3308	667.2752
3	$z_3 = y_3 - 0.8586x_3$	1.0573	103.5333
4	$z_4 = y_4 - 0.6574x_4$	0.6574	266.3676

根据上述的经验结果，本文中配对交易策略的开仓、平仓与止损阈值确定为：

1. 开仓规则：若第 t 日的价差组合 z_{it} 满足 $z_{it} > \mu_i + 0.75\sigma_i$ ，且此时仓位为 0，则在第 t 日开仓做空价差 z_i 。若 z_{it} 满足 $z_{it} < \mu_i - 0.75\sigma_i$ ，且此时仓位为 0，则在第 t 日开仓做多价差 z_i 。

2. 平仓规则：若第 t 日的价差组合 z_{it} 满足 $z_{it} \geq \mu_i$ ，且此前持有价差 z_{it} 的多头，则在第 t 日进行平仓。若 z_{it} 满足 $z_{it} \leq \mu_i$ ，且此前持有价差 z_{it} 的空头，则在第 t 日进行平仓。

3. 止损规则：若第 t 日的价差组合 z_{it} 满足 $z_{it} > \mu_i + 2\sigma_i$ ，且此前持有价差 z_{it} 的空头，则在第 t 日进行平仓止损。若第 t 日的价差组合 z_{it} 满足 $z_{it} < \mu_i - 2\sigma_i$ ，且此前持有价差 z_{it} 的多头，则在第 t 日进行平仓止损。

第4章 策略回测结果分析

4.1 交易成本参数设定

期货的交易成本主要包括保证金，手续费和滑点三个部分。其中不同期货的保证金和手续费比例并不相同，并且会由交易所不断进行调整。表4-1是上海期货交易所和大连期货交易所公布的2016年1月期货手续费和保证金比例，我们按照表中的数据对策略进行回测。

表4.1 各期货品种手续费

期货品种编号	期货品种	手续费	保证金
1	螺纹钢	万分之0.45	6%
2	焦炭	万分之0.6	5%
3	铝	3元/手	5%
4	锌	3元/手	6%
5	玉米	1.2元/手	5%
6	玉米淀粉	1.5元/手	5%
7	聚乙烯	2元/手	6%
8	聚丙烯	万分之0.7	6%

滑点是指期货交易过程中下单价格与成交价格不一致的现象，一般由交易延迟和价格的瞬时波动引起。由于本文的统计套利策略以日为频率进行交易，交易频率较低，因此受滑点的影响不大。本文在回测中的滑点假设为万分之2.5。

4.2 策略评价指标

本文主要采用年化收益率，年化波动率，夏普比和最大回撤四种指标来对投资策略进行评价。下面我们将对这四种指标一一进行解释。

1. 年化收益率。年化收益率是根据当前累计收益率计算的策略平均每年能够获得的收益率，设策略在累计 N 个交易日中获取了 R 的累计收益率，则年化收益率的计算公式为：

$$R_y = (1 + R)^{\frac{250}{N}} - 1. \quad (4-1)$$

2. 年化波动率。年化波动率是根据策略每日收益率的波动率推算出的策略年收益率波动率。设策略的每日波动率为 σ_d ，则 σ_d 等于策略每日收益率的标准差。而年化波动率 σ_y 的计算公式为：

$$\sigma_y = \sqrt{250}\sigma_d. \quad (4-2)$$

3. 夏普比(Sharpe Ratio)。夏普比是对投资策略的收益和风险进行综合评价的一个指标。一般来说，策略所承担的风险越大，收益率就越高。而夏普比衡量了策略每承担一单位风险，会产生多少超额收益。一般来说，夏普比是衡量基金业绩和投资策略有效性最常用的一个指标。夏普比(Sharpe Ratio)的计算公式为：

$$\text{Sharpe Ratio} = \frac{R_y - R_f}{\sigma_y}. \quad (4-3)$$

其中 R_y 和 σ_y 表示年化收益率和年化波动率，而 R_f 表示年化无风险利率。

4. 最大回撤(Max Drawdown)。最大回撤计算了策略在回测期中一段连续时间内取得的最大亏损，一般用来衡量使用该策略可能遇到的最坏情况，是一个非常重要的风险评估指标。设回测期共有 N 个交易日，对于 $0 \leq t_1 < t_2 \leq N$ ，设投资组合在第 t_1 天和第 t_2 天的总价值为 $P(t_1)$ 和 $P(t_2)$ 。则策略从第 t_1 日到第 t_2 日的亏损率为 $[P(t_1) - P(t_2)]/P(t_1)$ （等于收益率的相反数）。而最大回撤即是求策略回测期内可能取得的最大亏损率：

$$\text{Max Drawdown} = \max_{0 \leq t_1 < t_2 \leq N} \frac{P(t_1) - P(t_2)}{P(t_1)}. \quad (4-4)$$

4.3 策略回测结果分析

我们首先在样本期（2015年-2017年）内对四个期货组合分别进行回测，由于我们在策略构建中用到了2015年-2017年内的价格数据，因此在样本期内的回测结果会相对偏高，但仍能在一定程度上反映策略的收益情况。表4.2记录了4个期

货组合在样本期内的回测结果。四组套利组合在回测期内都取得了 15%-25% 的年化收益率，有着比较可观的收益率。但其中螺纹钢-焦炭、铝-锌、玉米-玉米淀粉三个组合的最大回撤都较高，其中玉米-玉米淀粉的最大回撤高达 27.22%，这说明单个套利组合存在收益不稳定，风险较大的特点。聚乙烯-聚丙烯组合的回撤率仅为 4.92%，是四个套利组合中风险最低的一组，其夏普比也在四个组合中最高。

表 4.2 2015-2017 年（样本期内）回测结果

套利组合	年化收益率	年化波动率	夏普比	最大回撤
螺纹钢-焦炭	24.661%	0.1676	1.2629	16.54%
铝-锌	18.214%	0.1970	0.7802	16.42%
玉米-玉米淀粉	20.064%	0.3115	0.6259	27.22%
聚乙烯-聚丙烯	20.325%	0.0990	1.5842	4.92%

接下来我们使用样本外的 2018 年数据对四个组合进行回测。与 2015-2017 年的回测结果进行对比可以发现，四个套利组合在样本期外的表现都明显不如样本期内。尽管如此，螺纹钢-焦炭、铝-锌两个组合仍然取得了较高的年化的收益率。其中螺纹钢-焦炭组合的年化收益率较高，波动率和最大回撤较低，是四个策略中夏普比最高的策略；而铝-锌组合虽然年化收益率较高，但波动率和最大回撤也是四个组合中最大的一个，因此风险较高，夏普比低。聚乙烯-聚丙烯组合的收益率不高，但波动率和最大回撤也较低，属于风险小收益小的组合。而玉米-玉米淀粉组合在 2018 年取得了负收益，这主要是因为 2018 年的玉米-玉米淀粉价差波动率要明显小于 2015-2017 年，因此 2018 年全年只有一次开仓交易机会，最终取得了负收益。这也说明了当市场环境变动时，原有的统计套利策略可能不再有效，需要根据新环境中的数据对原有策略进行改进。

表 4.3 2018 年（样本期外）回测结果

套利组合	年化收益率	年化波动率	夏普比	最大回撤
螺纹钢-焦炭	18.486%	0.1351	1.0491	8.53%
铝-锌	13.650%	0.2612	0.4790	33.06%
玉米-玉米淀粉	-1.703%	0.2102	-0.1566	14.91%
聚乙烯-聚丙烯	5.998%	0.0411	0.5382	2.64%

接下来，我们按照每个组合分配资金相同的方式将四个套利策略组合成一个混

合策略，并对混合策略进行了回测，图 4.1 和表 4.3 是混合策略在 2015-2017 年和 2018 年的收益曲线与评价指标。与单个套利组合相比，混合策略的收益率并没有非常显著的提高，但波动率和最大回撤要比大部分套利组合小。这是由于不同的套利组合之间相对独立，通过混合策略的方式能够将风险分化，从而降低策略的风险。混合策略在 2015-2017 年和 2018 年都有着大于 1 的夏普比，其中在 2018 年的夏普比要高于所有的独立期货组合。

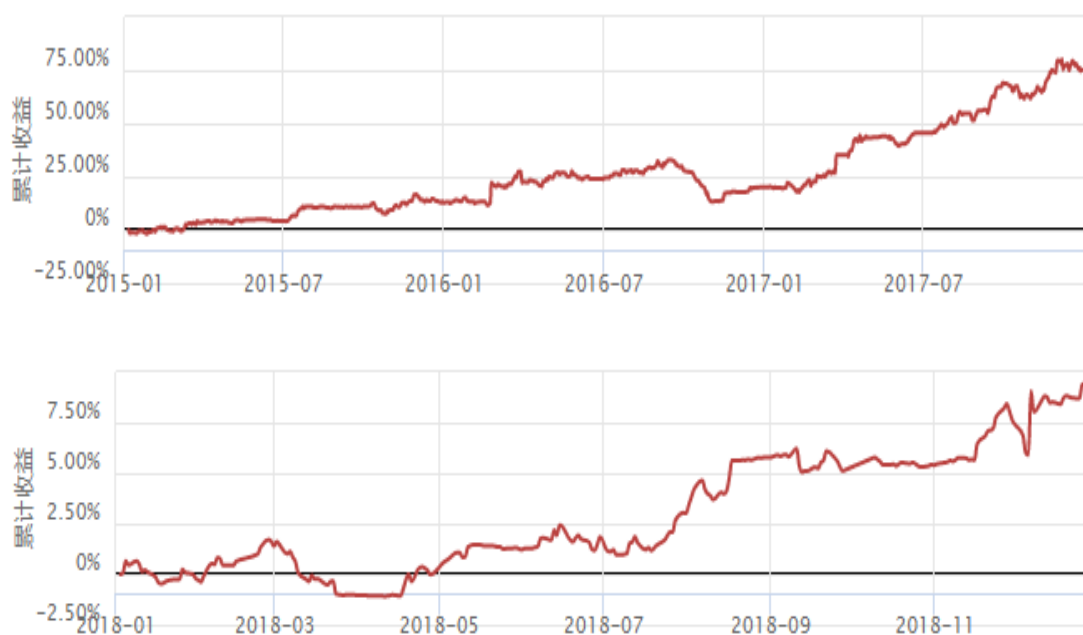


图 4.1 混合策略收益曲线图

表 4.4 混合策略回测结果

回测时间	年化收益率	年化波动率	夏普比	最大回撤
2015-2017	21.275%	0.1197	1.3951	12.87%
2018	9.399%	0.0481	1.1051	2.766%

第5章 结论

本文在中国商品期货市场上建立了基于协整模型的统计套利策略，并对策略进行了回测。根据商品期货的流动性与现实联系，本文选择了四个期货组合，并对各个组合进行了协整检验。四个期货组合都通过了协整检验，这说明两个期货价格序列的残差具有平稳性，可以用于配对交易。我们根据期货价格序列的回归系数确定了投资组合配比，并根据价差时间序列的均值和标准差确定了开仓与平仓阈值。经过回测，发现四个期货组合在样本期内都取得了 18%-25% 的收益率，但不同策略的风险情况差别较大，其中玉米-玉米淀粉组合的最大回撤为 27.22%，而聚乙烯-聚丙烯组合的最大回撤仅为 4.92%。四个投资组合在样本期外的表现都明显不如样本期内。其中螺纹钢-铁矿石组合的收益率和夏普比最高，收益率为 18.48%，夏普比为 1.05；而玉米-玉米淀粉组合在 2018 年的收益率为 -1.70%，未能取得正收益。这说明单个期货组合的收益表现不够稳定，风险依然较大。将四个期货组合按照等资金的方式组合为一个混合策略后，发现虽然混合策略的收益率没有提升，但其最大回撤和年化波动率有所下降，因此样本期外的混合策略夏普比为 1.11，高于所有的单个期货组合。这说明将独立性较好的不同策略进行组合能够分化部分风险，从而降低策略风险，提高混合策略的夏普比。

本文中的策略有两个可能的改进方向。一是选择更多有效的期货组合，通过这种方式来分化风险，提高混合策略的稳定性和夏普比。二是对价差时间序列进行波动率建模，并根据近期波动率的变化来调整开仓和平仓阈值。2018 年的玉米-玉米淀粉组合的价差均值与样本期内数据无太大变化，但价差波动率要明显小于样本期内数据，这导致了玉米-玉米淀粉组合未能在 2018 年获利。通过波动率建模能够根据价差波动率的变化来改变交易阈值，避免开仓阈值过高或过低导致的策略失效。

参考文献

- [1] Hogan S, Jarrow R, Teo M, et al. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. *Journal of Financial Economics*, 2004, 73(3):525-565.
- [2] Gatev E, Goetzmann W N, Rouwenhorst K G. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *Review of Financial Studies*, 2006, 19(3):797-827.
- [3] Do B, Faff R. Does Simple Pairs Trading Still Work? *Financial Analysts Journal*, 2010, 66(4):83-95.
- [4] Vidyamurthy G. Pairs Trading: quantitative methods and analysis. John Wiley & Sons, 2004.
- [5] Dunis C L, Laws J, Evans B. Trading futures spreads: an application of correlation and threshold filters. *Applied Financial Economics*, 2006, 16(12):903-914.
- [6] Elliott R, Hoek J V D, Malcolm W. Pairs trading. *Quantitative Finance*, 2005, 5(3):271-276.
- [7] Engle R F, Granger C W J. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 1987, 55(2):251-276.
- [8] Hull J. Options, Futures and Other Derivatives. Prentice Hall, 2011.
- [9] Zeng Z, Lee C G. Pairs trading: optimal thresholds and profitability. *Quantitative Finance*, 2014, 14(11):1881-1893.
- [10] Bertram W K. Analytic solutions for optimal statistical arbitrage trading. *Physica A Statistical Mechanics & Its Applications*, 2012, 389(11):2234-2243.
- [11] Cummins M, Bucca A. Quantitative spread trading on crude oil and refined products markets. *Quantitative Finance*, 2012, 12(12):1857-1875.
- [12] Do B, Faff R. Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 2012, 35(2):27.

致谢

衷心感谢我的导师唐宏岩教授对我的悉心指导。唐老师的风险理论课程让我对许多金融模型及其背后的逻辑有了更进一步的理解，对我启发很大。

感谢数学科学系各位老师对我的教导，你们的讲授让我学到了许多知识，为这篇论文的写作打下了坚实的基础。

最后，特别感谢我的父母。你们一直以来的关心、支持与鼓励为我的学习与生活带来了许多动力，谢谢你们！

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

