

基于随机过程与支持向量机 构建期货配对交易策略

刘 辉, 刘忠元, 周伟杰

摘 要: 运用协整分析法选取配对交易标的, 并通过随机过程(O-U过程)和支持向量机(SVM)预测价差的变化趋势, 构建一种新型的配对交易策略, 并与传统的配对交易策略进行了对比分析。选取2015—2017年郑州期货交易所(CZCE)菜粕期货与大连期货交易所(DCE)豆粕期货的一分钟交易数据进行实证分析, 结果表明该新型配对交易模型较好地预测了价差变化的趋势, 胜率和收益率明显优于传统配对交易策略。

关键词: 配对交易; 支持向量机; 预测; 收益率

作者简介: 刘辉, 理学博士, 常州大学商学院副教授、硕士生导师; 刘忠元, 常州大学商学院硕士研究生; 周伟杰, 管理学博士, 常州大学商学院讲师。

基金项目: 国家自然科学基金一般项目“分频灰色形态模型的构建及其应用研究”(71701024)。

中图分类号: F832.48 **文献标识码:** A **Doi:** 10.3969/j.issn.2095-042X.2018.03.007

配对交易策略起源于美国的华尔街, 是成熟资本市场的主流投资策略之一。该策略在美国股票市场一经推出, 便获得了巨大成功。随着国内做空机制逐渐放松, 基于统计套利的量化交易方式获得快速发展与应用。配对交易策略作为一种市场中性策略, 其思想主要是指从市场上找出一对历史价格走势相近的标的进行配对, 当配对标的之间的价差偏离历史均值时, 做空价格较高的标的, 做多价格较低的标的, 当价格回复到均值附近时, 结束头寸从而获得利润。

一、文献综述

关于配对交易, 国外学者的研究已经形成了一套成熟的理论体系。代表性的理论方法包括 Gatev 等^[1]提出的最小距离法、Vidyamurthy^[2]提出的协整分析法以及 Elliott 等^[3]提出的随机价差法。这些经典理论的提出为配对交易策略的实现提供了理论基础, 其在金融市场上的实际应用证明了配对交易策略的商业价值。

配对交易策略的相关研究主要包括配对标的的选取和交易参数的设计两个方面。在配对标的的选取方面, Gatev 等^[1]使用1962—2002年美国股票的日线收盘数据, 通过最小距离法筛选出合适的配对股票组合。王春峰等^[4]利用沪深300成分股2006—2009年的数据, 对基于最小距离法的经典配对交易策略进行了实证测算。在交易参数设计方面, Huck^[5]将神经网络方法和多属性决策理论相结合, 使用神经网络技术对候选配对的价差进行预测, 使用多属性决策技术对候选股票

进行排序;唐国强等^[6]利用切比雪夫不等式和夏普比率构建套利阈值统计量,研究了中国白糖期货合约数据的最优阈值以达到利润最大化。近年来,一些学者对配对交易方法进行评估比较。Bogomolov^[7]将距离法、协整法和随机价差法应用于澳大利亚证券交易所,发现这三种方法每年都能得到 5%~12% 的收益。

但相关研究集中于证券市场,且选取的交易数据大多为日线数据。随着信息透明度和市场有效性的提高,传统配对交易策略在国内金融市场的收益变得越来越低。本文以国内商品期货市场为研究对象,通过 O-U 随机过程与支持向量机预测价差的变化趋势,构建了一种新型的配对交易策略。选取 CZCE 菜粕与 DCE 豆粕期货进行跨市套利,使用 2015 年 9 月 8 日至 2017 年 9 月 8 日的 1 分钟高频交易数据进行实证分析,研究表明,该新型配对交易策略在国内期货市场中具有可行性,其胜率和收益率明显优于传统配对交易策略。

二、模型与方法

(一) 协整分析法

自 Engle 和 Granger^[8]提出了金融时间序列的协整理论和误差修正模型后,协整模型被广泛应用于时间序列建模。金融时间序列往往表现出非平稳性,而协整理论的贡献在于发现非平稳时间序列之间的线性关系,并进行线性组合为平稳序列。

本文采用 EG 两步法验证配对的两个金融时间序列之间是否存在协整关系。首先,单位根(ADF)检验。确定两个金融时间序列是否为同阶单整,如果是同阶单整,则进行最小二乘法(OLS)回归,计算出残差。然后,对残差进行 ADF 检验。如果残差平稳,则认为两个金融时间序列之间存在协整关系,否则认为两者不存在协整关系。

(二) O-U 模型

使用日线数据进行实证研究时,通常采用标准分数(z-score)对时间序列进行标准化。本文运用 1 分钟交易数据随机性较大,采用 z-score 标准化往往会导致偏差过大。因此,本文采用 O-U 过程描述价差序列均值回复的随机过程^[9]。此过程可描述为 $dX_t^{sp} = \alpha(u - X_t^{sp})dt + \epsilon dW_t$ ($\alpha > 0, \epsilon > 0, u \approx 0, W_t$ 表示维纳过程, X_t^{sp} 为两标的 t 时刻的价差)。经参数变换和伊藤转换,得:

$$X_{t+1}^{sp} = bX_t^{sp} + \sigma_{t+1}, \epsilon = \sqrt{\frac{D(\sigma)}{1-b^2}} \quad (1)$$

(三) 支持向量机模型

支持向量机是由 Vapnik 等人提出的一种分类算法,其基本模型是构建特征空间上的最大间隔线性分类器,通过寻求结构化风险最小来提高机器学习模型的泛化能力。线性分类器通过构建最优超平面将一组数据分为两类,这个线性超平面的一般形式 $f(x) = w^T x + b$, w^T 为权重, b 为阈值。为使分类超平面的几何间隔达到最大值,可求解以下目标函数的最优化解。

$$\begin{aligned} \max \quad & \frac{1}{\|w\|}; \\ \text{s. t.} \quad & y_i(w^T x_i + b) \geq 1, i = 1, \dots, n \end{aligned} \quad (2)$$

通过拉格朗日对偶性转化和 SMO 算法可求最优解 w^* 和 b^* 。

$$w^* = \sum_{i=1}^l \alpha_i^* y_i x_i; b^* = y_i - \sum_{i=1}^l \alpha_i^* y_i (x_i x_j) \quad (3)$$

(四) 技术指标

现有研究只考虑了两种标的之间的价差 (SPREAD) 因素, 忽略了交易过程的其他交易信息, 而这些信息也为配对交易提供有价值的信息。因此, 本文考虑了除价差外的其他多种技术指标差作为样本特征, 选取了算数移动平均 (SMA)、加权移动平均 (WMA)、相对强弱 (RSI) 和资金流量指数 (MFI) 等 4 个技术指标。

$$SMA_t = \frac{1}{n} \times \sum_{i=t-n+1}^t P_i^c \quad (4)$$

式中, SMA_t 为标的 t 时刻的 SMA 值, n 为时间跨度 (本文中 $n=5$), P_i^c 为标的 i 时刻的收盘价。

$$WMA_t = \sum_{i=t-n+1}^t W_i P_i^c, W_i = \frac{i - (t - n)}{1 + 2 + \dots + n} \quad (5)$$

式中, WMA_t 为标的 t 时刻的 WMA 值, W_i 为标的 i 时刻的权重。

$$RSI_t = 100 - \frac{100}{1 + RS_t} \quad (6)$$

式中, RSI_t 为标的 t 时刻的 RSI 值, RS_t 为标的 t 时刻前 n 个时刻内的涨跌幅均值比。

$$MFI_t = 100 - \frac{100}{1 + MFR_t}, MF_t = v \times \frac{P_t^h + P_t^l + P_t^c}{3} \quad (7)$$

式中, MFI_t 为标的 t 时刻的 MFI 值, MFR_t 为标的 t 时刻前 n 个时刻内的正负资金流量比, MF_t 为资金流量, v 为成交量, P_t^h 为最高价, P_t^l 为最低价, P_t^c 为收盘价。若 $\frac{P_t^h + P_t^l + P_t^c}{3} > \frac{P_{t-1}^h + P_{t-1}^l + P_{t-1}^c}{3}$, 则将 MF_t 记为 PMF_t , 反之记为 NMF_t , PMF_t 和 NMF_t 分别表示标的 t 时刻的正资金流量与负资金流量。

(五) 构建样本特征和标签

首先, 构建价差特征模型。

$$T_{sp} = \left| \frac{X_t^{sp} - u_{sp}}{\epsilon_{sp}} \right| \quad (8)$$

式中, T_{sp} 为价差特征值, X_t^{sp} 为价差, u_{sp} 为价差均值, ϵ_{sp} 为 O-U 过程计算的价差标准差。同理, 上述四种技术指标的差值可构建四类特征模型:

$$T_{sm} = \left| \frac{X_t^{sm} - u_{sm}}{\epsilon_{sm}} \right|, T_w = \left| \frac{X_t^{w} - u_w}{\epsilon_w} \right|, T_m = \left| \frac{X_t^m - u_m}{\epsilon_m} \right|, T_r = \left| \frac{X_t^r - u_r}{\epsilon_r} \right| \quad (9)$$

其次, 样本过滤。通过比较前后价差的变化幅度, 可判断每次交易是否获利, 以达到划分标签的目的。为了排除前后价差变化幅度微小不足以获利的样本, 根据以下算法对样本进行过滤。

$$m = |X_{t+1}^{sp}| - z \times |X_t^{sp}|, n = |X_{t+1}^{sp}| - (2 - z) \times |X_t^{sp}| \quad (10)$$

式中, X_{t+1}^{sp} 与 X_t^{sp} 分别表示标的 $t+1$ 与 t 时刻的价差, z 取 0.9。当 $m < 0$, 标签记为 “+1”; 当 $m \geq 0$ 且 $n \leq 0$, 标签记为 “0”; 当 $n > 0$, 标签记为 “-1”。排除标签为 “0” 的样本, 构建二分类的样本集。

(六) 构建 SVM 预测模型

选用 LIBSVM 软件包训练数据, 建立价差变化幅度预测模型。为了提高模型训练速度, 避免原始数据中部分特征范围过大而另一部分特征范围过小, 在建立训练模型之前需要对样本特征

规范化至 $[0, 1]$ 之间。模型参数 s 设置为“C-SVC”，核函数类型设为线性核函数，然后使用训练好的 SVM 模型对测试集进行预测。

(七) 构建交易信号

第一，开仓规则。选取 1 倍标准差作为开仓阈值，当价差标签为“+1”且价差偏离价差均值超过 1 倍标准差时，则执行开仓指令，即卖空相对被高估的标的，买入相对被低估的标的。

第二，平仓规则。开仓后当价差回复到均值附近时，则进行平仓操作。在本文中该平仓阈值选取 0.2 倍标准差。此外，由于本文选取的是 1 分钟期货的高频交易数据，不适合长期持有，因此设定当开仓后 3 个交易日内价差尚未到均值附近，则以第 3 个交易日的收盘价平仓头寸。

第三，止损规则。开仓后当价差继续偏离时，为了避免价差出现极端偏离而导致损失，需要设定止损线。本文设定止损线为 2 倍标准差，即当开仓后，价差继续偏离达到 2 倍标准差时，则执行平仓指令。

三、实证研究

(一) 数据选取

菜粕的蛋白含量约 36%，豆粕的蛋白含量约 43%。这两种饲料具备可替代性。正常情况下，菜粕和豆粕具有较为稳定的价差，这使得 CZCE 菜粕与 DCE 豆粕的价格联系更加紧密，为菜粕豆粕的跨市套利提供了可行性^[10]。本文选取 CZCE 菜粕与 DCE 豆粕期货 1 分钟高频交易数据进行实证研究，选取的时间区间为 2015 年 9 月 8 日至 2017 年 9 月 8 日，筛选出的样本点共计 181 720 个。

(二) 数据检验

1. 相关性分析

由图 1 可以直观地看出 CZCE 菜粕与 DCE 豆粕收盘价呈现趋同走势，两者的相关系数达到 0.938 2，说明两者之间具有很高的线性相关关系。

CZCE 菜粕与 DCE 豆粕收盘价的时间序列 X , Y 及其一阶差分序列的检验结果(见表 1)表明，原始收盘价序列的单位根检验 p 值远大于 0.000 1，接受原假设，即 CZCE 菜粕与 DCE 豆

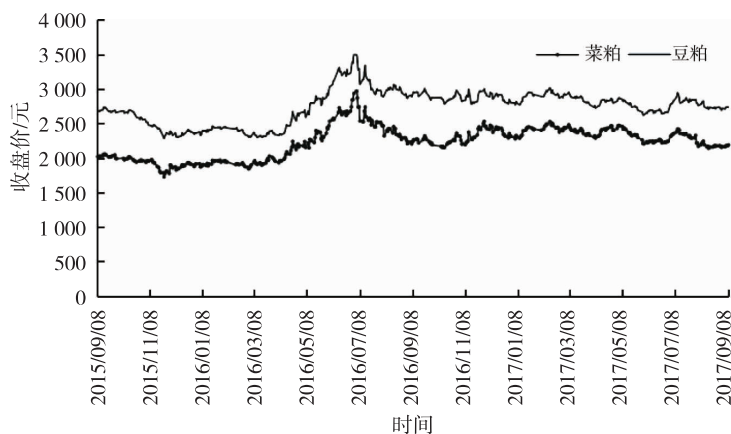


图 1 CZCE 菜粕与 DCE 豆粕收盘价走势图

粕的时间序列为非平稳序列。一阶差分序列 ADF 检验结果表明，两者的 p 值远小于 0.000 1，拒绝原假设，即一阶差分序列都为平稳序列。因此，CZCE 菜粕与 DCE 豆粕 1 分钟收盘价时间序列同为一阶单整。

2. 协整分析

根据 CZCE 菜粕与 DCE 豆粕 1 分钟收盘价建立协整回归方程，得到长期均衡关系 $X = 0.881 3 *$

$Y = 189.6428 + resid$ ($resid$ 表示残差项)。 $R^2 = 0.8803$, 表明该协整回归模型与样本数据的拟合度较高。对 $resid$ 进行 ADF 检验结果 (见表 2) 表明, 残差项不含有单位根, 残差序列为平稳序列, 该模型不存在伪回归现象。因此, CZCE 菜粕与 DCE 豆粕 1 分钟收盘价之间存在协整关系。

(三) SVM 预测结果分析

将 CZCE 菜粕与 DCE 豆粕 181 720 个样本数据进行标记筛选。标签 “+1” 与 “-1” 的样本数据基本均衡 (见表 3)。将数据集按 80%—20% 划分为训练集与测试集, 其中训练集的时间跨度为 2015 年 11 月 5 日 23 时 30 分至 2016 年 7 月 15 日 9 时 5 分, 测试集时间的跨度为 2016 年 7 月 15 日 9 时 6 分至 2017 年 9 月 7 日 14 时 31 分。

SVM 模型预测结果如表 4 所示。 $TP/(TP + FP)$ 表明该交易策略的胜率达到了 72.72%, $TN/(TN + FP)$ 表明该策略成功避免了 95.60% 的亏损交易。同时, 预测亏损与预测盈利的个数分别为 3 321 与 297, 说明该模型为风险厌恶型。同时, 该模型错过了 1 463 个交易获利机会, 产生了 81 个交易亏损机会, 说明此模型宁愿错过交易获利机会, 也要避免潜在的亏损交易。结合配对交易策略的交易信号, 测试集中 3 518 个样本点的套利时机如图 2 所示。

(四) 传统与新型配对交易策略比较

结合配对交易策略的套利规则, 分别对传统与新型配对交易策略进行比较分析,

模拟交易结果如表 5 所示。新型配对交易策略的胜率和收益率都大大优于传统配对交易策略, 这说明本文描述的新型配对交易策略, 提高了传统配对交易策略的胜率和收益率。

表 1 单位根检验

序列	ADF 值	p 值	原假设 *
菜粕 X	-2.070 7	0.548 9	接受
豆粕 Y	-1.976 2	0.589 4	接受
X 一阶差分	-475.11	<0.000 1	拒绝
Y 一阶差分	-464.62	<0.000 1	拒绝

注: 原假设 * 为序列非平稳。

表 2 残差单位根检验

序列	ADF 值	p 值	原假设 *
残差	-4.413 1	<0.000 1	拒绝

表 3 样本标签结构

	训练集	测试集	重构数据集
标签 “+1”	6 723	1 679	8 402
标签 “-1”	7 348	1 839	9 187
总计	14 701	3 518	17 589

表 4 SVM 模型预测结果分析表

	预测亏损	预测盈利	总计
真实亏损	$TN = 1\ 758$	$FP = 81$	1 839
真实盈利	$FN = 1\ 463$	$TP = 216$	1 679
总计	3 221	297	3 518

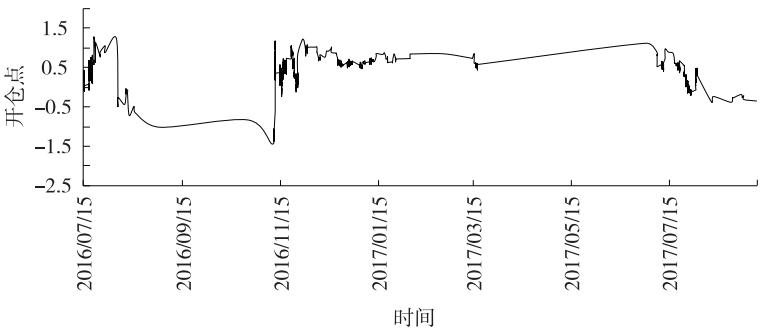


图 2 新型配对交易开平仓示意图

表 5 传统与新型配对交易策略的对比表

类型	套利次数	胜率/%	收益率/%
传统配对交易策略	64	57.81	10.70
新型配对交易策略	16	81.25	34.05

四、结论

本文设计了一种基于 O-U 过程和 SVM 优化的配对交易策略。以 2015 年 9 月 8 日至 2017 年 9 月 8 日时间段内 CZCE 菜粕与 DCE 豆粕期货 1 分钟高频交易数据为研究对象,检验了该配对交易策略的可行性。结果表明,引入 SVM 模型对传统配对交易策略进行优化,能较好地预测出价差变化趋势,从而适度战胜市场;新型配对交易策略的胜率明显高于传统配对交易策略,降低了配对交易的亏损风险,使新型配对交易策略的获利能力大大提升。新的模型为配对交易策略提供了新的思路,有助于改善传统配对交易策略的收益现状,提升配对交易策略的获利能力。为获取更好的交易表现,新模型仍需进一步地改进,例如在样本特征方面可以引入更多优异的技术指标,或者对技术指标做 PCA 分析;此外,在标签划分标准方面采取的是单一标准,未来研究可以尝试多种划分标准。

参考文献:

- [1] 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.
- [2] VIDYAMURTHY G. Pairs trading: quantitative methods and analysis [M]. Hoboken: Wiley, 2004: 35-47.
- [3] ELLIOTT R J, VANDER H J, MALCOLM W P. Pairs trading [J]. *Quantitative finance*, 2005, 5 (3): 271-276.
- [4] 王春峰,林碧波,朱琳. 基于股票价格差异的配对交易策略 [J]. *北京理工大学学报(社会科学版)*, 2013, 15 (1): 71-75.
- [5] HUCK N. Pairs selection and outranking: an application to the S&P 100 index [J]. *European journal of operational research*, 2009, 196 (2): 819-825.
- [6] 唐国强,高伟,覃良文,等. 基于切比雪夫不等式的白糖高频数据统计套利 [J]. *统计与决策*, 2016, 445 (1): 87-90.
- [7] BOGOMOLOV T. Pairs trading in the land down under [M]. Los Angeles: Social Science Electronic Publishing, 2010: 1-22.
- [8] ENGLE R F, GRANGER C W J. Co-integration and error-correction: representation, estimation and testing [J]. *Econometrica*, 1987, 55 (2): 251-276.
- [9] 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 (1): 3-11.
- [10] 周伟杰,顾荣宝. 股指期货和现货的线性、非线性 Granger 因果关系分析——基于 1 分钟高频数据的实证研究 [J]. *常州大学学报(社会科学版)*, 2015, 16 (4): 45-51.

A Pairs Trading Strategy Based on Stochastic Process and Support Vector Machine

Liu Hui, Liu Zhongyuan, Zhou Weijie

Abstract: By using cointegration analysis method to choose pairs of trading objects and stochastic process (O-U process) and Support Vector Machine to predict the trend of price changes, a new pairs trading strategy is proposed and compared with the traditional one. Through the empirical study of one-minute transaction data of rapeseed meal in Zhengzhou Commodity Exchange (ZCE) and soybean meal in Dalian Commodity Exchange (DCE) from 2015 to 2017, it demonstrates that the new pairs trading model well predicts the trend of price changes and both the accuracy rate and return rate are significantly better than the traditional pairs trading strategy.

Key words: pairs trading; SVM; predict; return rate

(收稿日期: 2017-10-27; 责任编辑: 沈秀)