基于高频数据的股指期货期现统计套利程序交易

张 连 华
(上海社会科学院数量经济研究中心 上海 200025)
(申银万国证券股份有限公司博士后科研工作站 上海 200031)

摘 要 我国最近推出股指期货后,大量投资者采用传统的基于持有成本理论的日间股指期货期现套利策略进场套利,使得期现套利的价差很快收窄,可套利机会越来越少。从两个角度对传统的股指期货期现套利策略进行拓展:一方面,将获取绝对收益的统计套利策略引入到股指期货期现套利中,并解决统计套利策略在进行股指期货期现套利时遇到的问题;另一方面,将交易标点的交易周期深入到分钟级别的高频行情。通过这两个方面的拓展,不仅开阔了股指期货期现套利的理论研究思路,而且获得了较好的收益风险系数,对投资实践也具有一定的指导意义。

关键词 统计套利 协整 程序交易 股指期货 期现套利

中图分类号 F830.91 **文献标识码** A

HIGH FREQUENT DATA BASED STOCK INDEX FUTURES PRESENT STATISTICAL ARBITRAGE PROGRAM TRADING

Zhang Lianhua
(*Research Center of Econometrics, Shanghai Academy of Social Sciences, Shanghai 200025, China*)
(*Shenyin & Wanguo Securities Co., Ltd., Shanghai 200031, China*)

Abstract After the Chinese government implements the stock index futures recently, a lot of investors follow the conventional cost holding theory based daytime stock index futures present arbitrage strategy to trade for profit, thus quickly narrowing down the price gap for present arbitrage and eliminating the opportunities for arbitrage. The paper extends the conventional stock index present arbitrage strategy from two angles. On the one hand, it introduces the statistical arbitrage strategy that obtains absolute gains into stock index futures present arbitrage, meanwhile solving problems that statistical arbitrage strategy encounters at stock index present arbitrage; on the other hand, it helps trading spot's trading cycle to penetrate into minute-level high frequent market quatation. By extending from these two angles, the paper not only broadens the stock index futures present arbitrage's theoretical research path, but also gains fine profit-risk efficient; additionally it is more or less instructive for investment practices.

Keywords Statistical arbitrage Co-integration Program trading Stock index futures Present arbitrage

0 引 言

中国资本市场推出股指期货对投资者的投资策略具有深远的影响。股指期货不仅可以被机构投资者用于股票等现货组合的套期保值,也可以用于资产配置和股票等权益市场的价格发现。其中,跨越股指期货市场和证券现货市场两个市场的股指期货期现套利对提高两个市场的流动性、增强两个市场的价格发现功能具有十分重要的意义。中国股指期货市场自 2011 年 4 月份推出后,许多投资者(特别是机构投资者)进入市场进行期现套利,使得期货市场和现货市场的联动和价格发现能力有明显提高。但是,目前从事股指期货期现套利的投资者具有两个共同特点:一是交易周期一般是以日为周期的日间交易为主;二是股指期货期现套利无套利区间的确定一般采用持有成本理论。在这个策略和理论的指导下,目前从事股指期货期现套利的投资者普遍认为,现阶段可操作股指期货期现套利机会明显减少,股指期货期现套利已经很难作为公司稳定的利润成长点之一进行投资。针对该问题,许多证券公司和基金公司研究机构对此进行了广泛研究,但是许多研究最终却是确证该论点。在长期跟踪研究国外高频交易和对冲基金策略的基础上,本文努力借鉴高频交易和基于协整的统计套利的思路,从股指期货期现套利交易执行的周期和交易赖以进行的理论两个角度对股指期货期现套利进行拓展,以尝试股指期货期现套利的模型并努力提高股指期货期现套利的实务效果。

高频数据具有和中低频数据不同的模式,交易换手率更高,可利用的套利机会更多,且一般在日内进行交易操作,相应的风险也主要集中在日内波动上。其次,统计套利是一种获取绝对收益的策略之一,可以交易的模型更广,是对持有成本理论的有益补充。现有的统计套利研究主要基于参数或非参数方法对价

收稿日期:2011 - 07 - 05。2011 年度上海市博士后科研资助计划重点项目(11R21421700)。张连华,博士,主研领域:高频与算法交易,高性能网络与安全。

差等标的进行建模并进而捕捉套利机会。国外文献对于统计套利的研究论文大部分集中在如何建立统计套利模型实现交易策略方面。文献[10,13]采用基于标准差距离度量的非参数化方法进行统计套利建模。文献[5]利用协整技术构造国际股票指数 FTSE 的一个组合来计算 FTSE 收益,组合权重由协整回归系数给出。文献[4,12]将协整技术运用于主动资产配置,并使用协整技术实现增强型指数追踪。利用该增强型指数追踪技术建立两个自制指数资金组合的统计套利交易策略。文献[8]首先提出采用状态空间方法对多空资产(或者组合)统计套利的价差进行建模。文献[7]提出在随机价差过程对应的收益水平差(文章称为随机残余价差)基础上建立均值恢复过程模型。与非参数化方法紧密相关的神经网络、遗传算法、机器学习和数据挖掘方法等目前也在统计套利中获得了应用[6,11,14,15]。国外的统计套利模型很多用在股票的配对套利上或者是不同债券的统计套利上。由于股指期货市场和融资融券市场的最近发展,国内很多研究人员采用协整等统计套利模型研究中国股票配对[1]和股指期货的跨期套利[2]。

本文将协整模型用于股指期货期现统计套利的高频交易信号选择中,通过尽量捕捉较大的套利机会来增强收益,减小风险。其他统计套利模型也可以用于股指期货期现套利,本文仅以协整模型为例子,给出交易框架,实务中投资者可以采用不同的统计套利模型以免出现交易同质化风险。

1 高频协整股指期货期现统计套利程序交易算法

先阐述所采用的股指期货期现统计套利的现货构建方法;然后基于协整统计套利基本原理;最后给出程序交易的基础信号算法,即高频协整统计套利交易算法。

1.1 股指期货期现统计套利的现货构建

与传统基于持有成本理论的股指期货期现套利模型类似,本文选择股指期货和现货间进行高频协整统计套利。由于目前没有股指期货标的 HS300 股票指数对应的交易产品(除了全部成分股),现货一般采用复制技术进行替代,ETF 产品组合在交易成本等方面具有显著优势[3],本文采用流动性高的 1 分钟高频 ETF 产品组合进行现货替代。

本文使用跟踪误差最小化技术进行 ETF 基金产品组合构造现货组合。跟踪误差,是指拟合标的指数的投资组合收益率与目标指数收益率之间的偏差。从理论上讲,投资组合可以完全跟踪目标指数的收益率。但在实际操作中,由于跟踪投资组合必须面对现实市场中各种不同类型的投资规则和约束条件,因此与目标指数之间往往存在一定的差异。造成跟踪误差的原因主要是成份股公司行为(分红、配股、增发、可转换债券转股票以及流通股本的回购)、目标指数对成份股的调整、资金配置过程中的四舍五入计算原则,或者基金经理的操作误差。

为了方便对不同组合间指数跟踪的精度的比较,定义统一的考察误差项的变量:

$R_{I,t}=\ln(I_t/I_{t-1})$ 沪深 300 指数第 t 分钟分钟对数收益率,I_t为指数分钟收盘价。

$R_{i,t}=\ln(p_{i,t}/p_{i,t-1})$ 组合中第 i 只股票/基金第 t 分钟分钟对数收益率。

$TE_t=R_{I,t}-\sum_{i=1}^{n}w_i\cdot R_{i,t}$ 分钟收益率偏差。

$ATE=\frac{1}{T}\sqrt{\sum_{t=1}^{T}[R_{I,t}-\sum_{i=1}^{n}w_i\cdot R_{i,t}]^2}$ 组合分钟均跟踪误差。

跟踪指数的最优化目标函数如下:

$$\text{Minimize } ATE=\frac{1}{T}\sqrt{\sum_{t=1}^{T}[R_{I,t}-\sum_{i=1}^{n}w_i\cdot R_{i,t}]^2}$$

$$\text{Subject to } \sum_{i=1}^{n}w_i=1\quad w_i\geqslant 0$$

1.2 基于协整的统计套利基本原理

基于协整的统计套利基本原理如下:首先,构造由多头头寸和空头头寸组成的复合资产组合,利用协整检验量检验动态价格或收益的预测能力;其次,构造协整回归,建立误差修正机制;最后,实施交易系统,开发资产收益可预测的成分。

变量间存在协整关系的前提条件是非平稳的变量序列具有相同的单整阶数。平稳性检验可通过 ADF 检验实现。同阶单整序列的协整检验和误差修正模型可通过 Engle-Granger 两步法实现[9]。

1.3 高频协整统计套利程序交易算法

假设沪深 300 股指期货和沪深 300 指数每个交易日 1 分钟价格序列为$\{IF_t,1\leqslant t\leqslant T\}$和$\{I_t,1\leqslant t\leqslant T\}$,记其对数价格序列为$\{\ln(IF_t),1\leqslant t\leqslant T\}$和$\{\ln(I_t),1\leqslant t\leqslant T\}$,并设已经计算得到追踪指数的 *ETF* 权重为$\{w_i,1\leqslant i\leqslant n\}$。算法步骤如下:

(1) 选取对数价格序列前面 T_1 个数据进行训练,得到一些交易参数。首先利用 Engle-Granger 两步法对 T_1 个数据进行协整判定:若是不协整,算法结束;若是协整,转(2)。

(2) 考虑不带常数的一元回归模型$\ln(IF_t)=\beta^*\ln(I_t)+\varepsilon_t,1\leqslant t\leqslant T_1$,利用 OLS 估计出$\beta$,计算价差$spread_t=\ln(IF_t)-\beta\cdot\ln(I_t),1\leqslant t\leqslant T_1$,计算其均值$\mu$和标准差$\sigma$。

(3) 设定参数λ、δ和χ。开始循环,$t\leftarrow T_1+1$。用 open 代表持仓状态,open = 0 为空仓;open = 1 代表持仓。

(4) 若$t=T$,如果 open = 1,强制平仓,open ← 0;若$t<T$,转(5)。

(5) 计算$spread_t=\ln(IF_t)-\beta\cdot\ln(I_t)$。

如果 open = 0:若$\mu+\lambda\cdot\sigma<spread_t\leqslant\mu+\chi\cdot\sigma$(记这种情况为 A 类),卖空 1 单位股指期货,同时将$\beta\times I_t\times 300$的资金按照$\{w_i,1\leqslant i\leqslant n\}$权重分别投资于各 ETF,open ← 1。

若$\mu-\chi\cdot\sigma\leqslant spread_t<\mu-\lambda\cdot\sigma$(记这种情况为 B 类),买入 1 单位股指期货,同时将$\beta\times I_t\times 300$的资金按照$\{w_i,1\leqslant i\leqslant n\}$的权重对应分配于 ETF,将分配资金用于融券做空 ETF,open←1。

如果 $open=1$:若是 A 类开仓,如果$spread_t\leqslant\mu-\delta\times\sigma$,平仓止盈,$open$←0。

若是 A 类开仓,如果$spread_t>\mu+\chi\cdot\sigma$,平仓止损,$open$←0。跳出循环,停止交易。

若是 B 类开仓,如果$spread_t\geqslant\mu+\delta\times\sigma$,平仓止盈,$open$←0。

若是 B 类开仓,如果$spread_t<\mu-\chi\cdot\sigma$,平仓止损,$open$←0。跳出循环,停止交易。

$t\leftarrow t+1$,转(4)。

2　高频协整股指期货期现统计套利算法的实验数据与结果分析

2.1　数据获取

为了保证算法的实效性,两种模型所需要的 1 分钟高频数据全部按照与未来实时交易时一致的环境实时获得。图 1 表示了行情高频研究数据的获取过程。

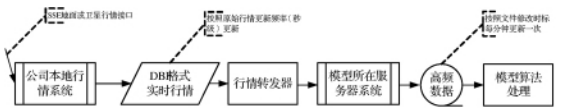

图 1　高频行情数据的获取过程

首先通过证券公司本地行情系统将全部行情数据按照上海证券交易所行情接口每 3 秒一次实时接收下来,并将其临时保存到 DBF 格式行情数据库中,然后通过转发器将此实时数据库转发到模型所在服务器系统,模型所在服务器系统通过识别文件更新标记按照每分钟一次的频率对研究股票池中的实时行情数据进行采样,将获得的 1 分钟原始高频数据按照时间顺序保存到 Excel 文件中作为模型的输入文本,并转入模型进行算法处理。

根据前述算法,整个交易数据可以分为三类:

(1) 1 分钟数据 ETF,目前主要研究上证 50 指数、上证 180 指数、深圳 100 指数三个流动性靠前的 ETF,并根据前述跟踪误差算法算出三只 ETF 替代 HS300 现货成分股的权重。(2) 沪深 300 指数 1 分钟数据。(3) 股指期货 1 分钟数据,包含自上市以来的所有 1 分钟高频数据,每期有 4 个合约,本文实证时选取 IF1005 进行研究。

表 1 给出了两个模型所采用的共同参数。

表 1　两个算法采用的共同参数

成本	股指期货	单边 0.1%	
	ETF	单边 0.1%	
样本总数	6134		
训练样本	1000		
权重	50ETF	100ETF	180ETF
	0.505	0.334	0.161
合约	IF1005		

2.2　高频协整统计套利程序交易算法的实验结果

借助获取的高频行情数据,将其输入到实时交易模型中,对"高频协整股指期货期现统计套利程序交易算法"进行测试和实证分析。表 2 给出了基准模型的实证结果统计,即高频协整统计套利程序交易算法的交易结果。

表 2　高频协整统计套利程序交易算法的交易结果统计

累计收益率(一个月)	平均每笔收益率	交易次数	交易分笔收益率								
3.57%	0.40%	9	-0.08%	1.29%	-0.34%	-0.49%	5.40%	-0.30%	1.02%	-2.81%	-0.12%

图 2 是高频协整统计套利程序交易算法的交易结果输出图形,从图中可以清楚地看出每次开仓、平仓(包括止盈、止损)的信号发生情况,和开平仓阀值的设置情况。图中纵向线表示训练期和交易测试期的分割线,分割线前为训练期,从分割线开始为交易测试期。

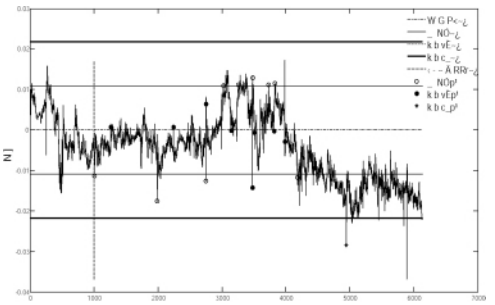

图 2　高频传统协整统计套利程序交易算法的实证结果图形

表 2 和图 2 的交易情况是相对应的。可以看出,基于协整模型的交易结果比目前业界采用的持有成本模型[3] 有较好的交易结果。与合理设计(指无套利交易边界的设计要综合考虑各种实际成本)的持有成本模型不同,基于协整模型的股指期现套利交易收益可能有负值,但是套利交易次数也略高。由于每次发生亏损的交易亏损较小,从而导致整体收益比持有成本模型有显著改善。

2.3　实验结果分析

下面进一步分析上节"高频协整股指期货期现统计套利程序交易算法"的交易结果相对传统的持有成本模型有所改善的原因。首先,"高频协整股指期货期现统计套利程序交易算法"采用更高的频率导致了更多的交易次数(传统的持有成本模型在实验期仅有 3 次,而高频协整股指期货期现统计套利程序交易算法有 9 次),这主要是由于两个因素导致的:一方面低频数据过滤了日内的价格波动,这些过滤了的日内价格波动不能产生股指期货期现套利交易信号;另一方面,高频条件下的交易波动率、价格演化周期等特征和低频不同,其影响因素不能仅仅为持有成本模型所解释。其次,统计套利模型具有灵活的参数调整方式,只要套利收益超过套利交易的成本,一旦有信号发生,就可以执行套利,该策略可以捕获到持有成本模型不能解释的其他因素导致的套利信号。当然,统计套利也有弱点,如果统计套利自身模型存在风险,交易可能会出现负收益,但是通过合适设计的统计套利模型和风险控制手段,这些风险是可以控制的。最后,从实验结果来看,如果套利时点持续存在时间越长,算法的效果越显著。

3　结论与建议

本文结合高频交易策略和基于协整的统计套利策略于股指期货的期现套利交易系统设计和实现中,阐述了股指期货期现套利的 ETF 现货构建方法,说明了基于协整的统计套利基本原理,给出了基于高频数据和协整理论的股指期货期现统计套利交易算法。高频数据具有和中低频数据不同的模式,交易换手率更高,可利用的套利机会更多,且一般在日内进行交易操作,相应的风险也主要集中在日内波动上。统计套利是一种获取绝对收益的策略之一,可以交易的模型更广,是对持有成本理论的有益补充。文章通过实验测试表明,新的算法不仅增加了交易次数,提高了系统总体收益率,而且整个交易系统的收益风险比率相对传统持有成本模型有显著提高。　　(下转第 156 页)

(1) 用户界面组件

本系统首先要设计一个客户程序,即人机界面,使用户可以对模型进行交互式操作及设置相应参数,然后调用相应组件来实现系统的各个功能。

(2) VRML 模型加载与显示组件

VRML 模型加载与显示组件包括模型数据结构的建立、模型视图的基本操作等功能,具体实现细节在文献[8]中进行了讨论,该组件打开.wrl 的文件,之后仍保存为.wrl 的文件,用于分层处理组件的输入。

(3) 模型分层处理组件

该组件以 VRML 文件为输入,进行分层处理后,以 CLI 层片文件格式存储。用于 RP 系统分层处理后续模块的处理。

4 应用实例与结论

本文利用 VC++和 OpenGL 在 Windows2000 环境下开发了将 VRML 文件分层输出为 CLI 文件的分层软件系统。图 3 为一个应用实例。在多面体表示上,VRML 采用顶点索引技术大大减少了 STL 存在的数据冗余。与 STL 相比,VRML 可以用很少的数据精确表示基本几何体及其变换,比如一个直径为 100 mm 的球体,给定模型误差 0.2 mm,STL 需要用约 20 000 个三角形来表示,而 VRML 只用一个半径值和圆心表示;对基本几何体而言,用几何法直接分层,分层精度高、速度快。

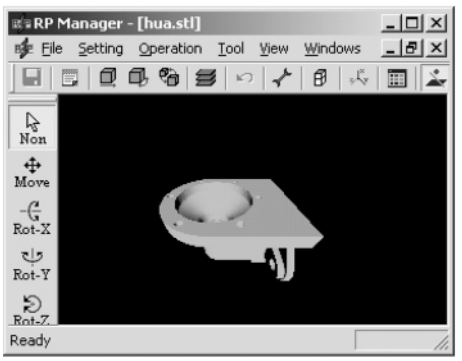

图 3 分层应用实例

VRML 是 Internet 上传输 VR 图像的规范,用户通过 Web 浏览器可以方便地观察 VRML 模型,这样通过 Internet 在公司内部或与用户之间建立了有效的通信交流,因此基于 VRML 的 RP 系统适应远程虚拟制造和 Color RP 的发展趋势。为了进一步提高成型精度,有必要研究支持曲线和彩色信息的新层片文件格式。

参 考 文 献

[1] Petitjean S. A survey of methods for recovering quadrics in triangle meshes [J]. ACM Comput. Surv,2002,34(2):211-262.

[2] Ling Waiming,Ian Gibson. Specification of VRML in Color Rapid Prototyping [J]. International Journal of CAD/CAM,2002,1(1):1-9.

[3] Korean Society of Precision Engineering. A Study on Rapid Prototyping using VRML Model [J]. International Journal of the Korean Society of Precision Engineering,2002,3(02):5-14.

[4] Li Zhanli,Sun Xiuying. Computing the intersection of a plane and geometric primitives in VRML model for rapid prototyping software [J]. Wuhan Ligong Daxue Xuebao,2006,8(1):918-923.

[5] VRML. The Virtual Reality Modeling Language [S]. International Standard ISO/IEC 14772-1:1997.

[6] 李占利,孙秀英. 一种实现 VRML 中坐标变换的方法 [J]. 西安科技大学学报,2006,26(2):240-244.

[7] 覃铭坚. 基于组件技术的快速成型软件系统开发 [D]. 西安:西安科技大学,2007.

[8] 李占利,孙秀英. RP 软件中 VRML 模型的可视化研究 [J]. 计算机工程与设计,2007,28(5):1185-1188.

[9] Mani K,Kulkarni P,Dutta D. Region-based adaptive slicing [J]. Computer-Aided Design,1999,31(5):317-333.

[10] Mund G B,Mall R,Sarkar S. An efficient dynamic program slicing technique [J]. Information and Software Technology,2002,44.

(上接第 95 页)

参 考 文 献

[1] 陈守东,韩广哲. 统计套利模型研究—基于上证 50 指数成份股的检验 [J]. 数理统计与管理,2007(26):5.

[2] 仇中群,程希骏. 基于协整的股指期货跨期套利策略模型 [J]. 系统工程,2008,26(12).

[3] 湘财证券. 股指期货期现套利解决方案 [R]. 2010-03.

[4] Alexander Carol. Optimal hedging using cointegration [J]. Philosophical Transactions of the Royal Society of London, Series A-Mathematical Physical and Engineering Sciences,1999,357:2039-2058.

[5] Burgess A N,Refenes A N. 1996a, Modelling non-linear cointegration in international equity index futures [C]//Neural Networks in Financial Engineering. World Scientific, Singapore,1996:50-63.

[6] Burgess A N. Statistical yield curve arbitrage in eurodollar futures using neural networks [C]//Neural Networks in Financial Engineering. World Scientific, Singapore,1996:98-110.

[7] Do B,Faff R,Hamza K. A new approach to modeling and estimation for pairs trading [R]. working paper (Monash University),2006.

[8] Elliott R,Van J. der Hoe,W. Malcolm. Pairs Trading [J]. Quantitative Finance,2005,5(3):271-276.

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

[10] Gatev E,Goetzmann W N,Rouwenhorst K G. Pairs Trading: Performance of a Relative Value Arbitrage Rule [R/OL]. Working Paper, Yale School of Management,1999. http://ssrn.com/abstract=141615.

[11] Giovanni Montana, Kostas Triantafyllopoulos, Theodoros Tsagaris. Flexible least squares for temporal data mining and statistical arbitrage [J]. Expert Syst. Appl,2009,36(2):2819-2830.

[12] Lucas A. Strategic and tactical asset allocation and the effect of long-run equilibrium relations [DB]. Serie Research Memoranda 0042, Free University Amsterdam, Faculty of Economics, Business Administration and Econometrics. De Boelelaan 1105, 1081 HV Amsterdam,1997.

[13] Nath p. High frequency pairs trading with US treasury securities: risks and rewards for hedge funds [R]. working paper, London Business School,2003.

[14] Philip Saks, Dietmar G. Maringer: Genetic Programming in Statistical Arbitrage [C]//EvoWorkshops,2008:73-82.

[15] Xing Fu, Avinash Patra. Machine Learning In Statistical Arbitrage [OL]. working paper,2009. http://www.stanford.edu/class/cs229/projects2009.html.