清 华 大 学

综 合 论 文 训 练

题目：配对交易——相对价值套利的算法实现与实验分析

培养单位：交叉信息研究院

学　　科：计算机科学与技术（计算机科学实验班）

姓　　名：陈炜艺

指导教师：唐平中教授

2014 年 6 月 20 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：陈炜亮　导师签名：唐平中　日　期：2014.6.20

中文摘要

论文分析一种统计套利策略——配对交易，即基于历史价格走势相似的成对股票存在长期均衡的假设，当它们之间的相对价值发散足够大时，卖空走势偏高同时买入走势偏低的股票，在收敛期进行套利。我们将采用“正交距回归”的残差定义发散的偏离程度，以度量相对定价的误差。任何长期均衡的偏离都会在后续的时间序列中收敛，基于此假设，配对交易策略的核心价值就是长期均衡附近的不断震荡。

在本论文中我们首先介绍基于协整的计量经济理论模型，再介绍配对交易策略核心的“误差纠正”性质，最后用例子和后台测试报告的形式构造一个最高历史收益的配对交易策略，实现过程引入机器学习算法进行实验。程序语言为 Python。

研究过程分两阶段，前半部分进行与配对交易策略相关的测试工具的开发和搭建，主要工程包括

- 定量软件工具包的开发环境搭建
- 投资组合评估与优化库
- 事件分析库
- 市场模拟库
- 时间分析库与市场模拟库的连接

后半部分进行配对交易的理论研究与实验探究，主要工程包括——

- 财务指标库
- 配对交易数学模型的理论构建与算法实现
- 后台库对配对交易策略的实验
- 构造最佳配对交易策略，生成最优决策报告

关键词：统计套利，配对交易，正交距回归，协整，误差纠正，机器学习

ABSTRACT

In this paper, we will introduce a statistical pairs trading strategy that is a relative value arbitrage on two equities and based on the premise there is a long-run equilibrium between the prices of the stock composing the pairs. We will use orthogonal distance regression to define the degree of deviation from the long-run equilibrium and represent the extent of mutual mis-pricing. Any deviation from the long-run equilibrium is compensated for in subsequent movements of the time series and this pairs trading involves trading on the oscillations about the equilibrium value.

We will give example step by step and experimental backtest report of this strategy after introducing the theoretical model based on the econometric paradigm of cointegration and error correction central to the analysis of this pairs-trading strategy. We will also introduce machine learning techniques to help improve the pairs-trading strategy. Programming language is Python.

Our research process can be divided into two parts, the former part is developer tasks accompanying with its analysis on pairs trading, as of -

- QSTK (Quantitative Software Toolkit) installation
- Portfolio assessment and optimization library
- Event studies library
- Market simulator library
- Event study library into simulator library

The latter part involves with quantitative analysis tasks as of -

- Financial features library
- Theoretical model and algorithm implementation of pairs trading
- Experiments of developed libraries on pairs trading strategy
- Construct highest Sharpe ratio strategy and generate backtest report

Key words: Statistical arbitrage pairs trading orthogonal distance regression cointegration error correction machine learning

目 录

第 1 章 引言 1
1.1 研究背景 1
1.2 论文目的 1
1.3 本论文配对交易策略的涉及领域 2
1.3.1 计算机算法辅助模型构造 2
1.3.2 代码实现 3
1.3.3 实验分析与金融应用 3
1.4 配对交易理论模型框架 4
第 2 章 配对交易理论模型 5
2.1 模型假设及目的 5
2.2 协整（相关程度） 6
2.3 剩余利差（事件定义） 9
2.4 交易例子 10
第 3 章 实现配对交易模型的算法 12
3.1 寻找并选择配对 12
3.2 后台测试 14
3.3 参数检验和对比 15
第 4 章 讨论和下一步研究 18
4.1 讨论 18
4.2 进一步研究 18
第 5 章 Python 代码库的实现 20
5.1 建立 QSTK 开发环境 20
5.2 投资组合的评估和优化 20
5.2.1 评估优化库的调用方法 20
5.2.2 评估优化库的实现 22

5.3 事件分析库 25
5.3.1 事件分析库的调用方法 25
5.3.2 事件分析库的实现方法 26
5.4 市场模拟库 28
5.4.1 市场模拟库的调用方法 28
5.4.2 市场模拟库的实现方法 29
5.5 连接事件分析库和市场模拟库 32
5.5.1 连接库的方法 32
5.6 实现金融预测指数（机器学习特征量） 33
5.6.1 金融预测指数的调用方法 33
5.6.2 用布林带进行事件分析 35
5.7 Indicators 库 37
插图索引 38
表格索引 39
公式索引 40
参考文献 41
致 谢 42
声 明 43
附录 A 外文资料阅读报告 44
A.1 Principle 44
A.1.1 Cointegration (correlation) 45
A.1.2 Residual spread (event definition) 46
附录 B 外文资料的调研 48
B.1 Pairs trading 48
B.2 Losing value 49
B.3 Market news analysis 50
B.4 Conclusion 50

第 1 章　引言

1.1　研究背景

19 世界 80 年代，华尔街量化分析师 Nunzio Tartaglia 建立了一个组织，该组织囊括了各名校的物理、数学和计算机科学研究员，目的在于挖掘资本市场中的套利机会。Tartaglia 的组织采用了当时学术导向用的复杂统计模型发展高科技交易项目，这个项目在当时就能通过自动交易系统运行。自动交易系统在当时是首创，一度震惊华尔街，原因在于它某种程度上替代了交易员（Trader）的职能，最关键的是它量化了人为策略，有计划并且恒久快捷的交易都让人认同它的巨大潜力。

除此之外，Tartaglia 也在自动交易系统的应用中发现了某些证券之间的价格同步性，即倾向于向同一形势发展。它们通过买入卖空在 1987 年就创造了极大的财富——一年在该策略给它们公司摩根斯坦利（Morgan Stanley）带来的收益达到 5 千万美元。虽然原摩根斯坦利在之后两年业绩不佳，在 1989 年分拆成了 JP Morgan 和 Morgan Stanley，配对交易（pairs trading）从此不再是该公司的私有策略，并开始在“市场中性”（Market Neutral）投资策略领域风靡，许多个人投资者和对冲基金公司也不断改进和进行该策略的开发与运营。

众所周知，任何套利策略的风靡和知名都会导致它的回报不断减小。在纽约时报（New York Times）的一篇报道中，过去 Tartaglia 手下的一个软件开发师 David Shaw，现已经是一间相当成功的量化组织 D.E. Shaw 的总裁提及，量化套利策略的利润现已经缩水，他的公司之所以能成功，主要因素还是在于较早使用该技术进入市场。Tartaglia 更以一种哲学性的口吻描述他的配对策略，他说人们一般喜欢在股票涨了以后才买入，而不愿意去买跌了的股票。换言之就是发现了策略好用后才加以利用，但为时已晚。那到底现在配对交易者是否还能从满是过激的无规律个人投资者中有规律地盈利？这是本论文的目的所在。

1.2　论文目的

综上背景，华尔街出现了量化套利（quantitative arbitrage）并长期成为风靡金融数学世界的交易方法。其中一种短期的套利策略被称之配对交易（pairs

trading)。该策略在华尔街风靡至少 20 年，并且现在仍是对冲基金和投资银行最主流的统计套利（statistical arbitrage）工具之一。配对交易实质概念比较简单，简例言之，如果能找到两只历史价格走势相似的股票，当它们之间的传布（spread）足够宽时，卖空走势偏高的股票，同时买入走势偏低的股票。如果未来重复它们价格走势相似的历史特性，相对价格将会收敛，之前的套利策略就能盈利。该策略仅仅是基于历史价格动态和简单的逆势原理（contrarian principle），难以让人相信它真的可以赚钱。如果美国资本市场在任意时刻都是充分有效的，配对交易在风险调整后是难以有回报的，当然真正的市场无法达到任意时刻充分有效，因此套利策略的优越在于实验模型的精益求精，该论文的目的正是在引入计算机算法后进行不断实验构造一个最高历史收益性的配对交易策略。同时也将回答背景中提出的问题“配对交易者是否还能从满是过激的无规律个人投资者中有规律地盈利?”，证明现在的逆势回报有一部分实质来自于人们对公司新闻信息的过激反应，而非价格回归模型中假设的长期相对均衡。

1.3 本论文配对交易策略的涉及领域

该论文中，我将对配对交易的风险和回报特性进行研究，数据为 2004 至 2013 年的美国股市日数据。由计算机算法、代码实现细节、金融应用三个层面进行介绍——

1.3.1 计算机算法辅助模型构造

首先会根据著名书籍《量化方法与分析》（Quantitative method and analysis）介绍的经典配对交易策略模拟构架，建造三步模型：

- 数据挖掘（Data Mining）算法进行高协整（Co-integration[1]）配对的搜索
- 机器学习（Machine Learning）算法进行高盈利性（Profitability）配对的筛选
- 后台测试（Back test）策略，生成金融交易报告

初期采用简单的算法进行配对的发现、选择和交易，研究几种直接、自融资交易法则的盈利效果。根据之前的简单实验，我发现对于好的配对股票组合，年收益大致在 11%。虽然配对策略只是挖掘股票信息的临时成分，但可以理论证明利润并不只是像文献所述的仅有均值回归，对此也将该论文算法层面陈述。

1.3.2 代码实现

该论文的实验实现以 Python 为主，代码将兼顾风险因子，以便合理研究结果的稳定性，即不只包含广泛使用的因子如价格数据，同时也兼顾低频的机构因素如破产风险，均加以量化。另外，在后台模拟测试（back test）代码中也会考虑金融微观因素如买入与卖出滑移（slippage），卖空利率（short-selling interest）和交易成本（transaction cost），更趋近真实交易。

1.3.3 实验分析与金融应用

考虑现实因素会降低超额收益的大小，但在初步实验看来，配对交易依然我的样本数据和改进模型中依然保持着盈利性，精确地说，在近十年的后台测试中有八年都维持着正回报。在基于以上算法的模型构造和代码实现后，我将进入实验分析阶段进行各模型的搭配模拟测试效果，包括改造回归模型，多种数据挖掘算法对配对进行搜索选择，多种机器学习算法对高利润型配对进行筛选，以及对量化因素可靠性的时间序列分析，最终以金融分析中的夏普比率（Sharpe ratio）进行策略优劣的比较。完成该部分试验后，将在论文中介绍收益性效果最好（即夏普比率最高）的算法搭配组合和策略模型，展现金融分析报告。

图 1.1　通过线性回归方法找到的一组配对：我们会在配对的回报率发散（价格走势相离）后打开头寸，而在回报率收敛后关闭头寸。（图片截自 Google Finance）

1.4 配对交易理论模型框架

为了构造一个市场中性投资组合，即同时构建多头和空头头寸以对冲市场风险，在任何市场环境下均能获得稳定收益的策略，我们发现两只历史价格相关系数很高的配对资产在未来依然保持着很强的相关系数。下图展示了一个均值回归配对的典型例子。通过他们的相关性，我们可以用他们的回报率数据进行线性回归分析，然后分析他们之间的相对定价与预期是否有一个明显的分离。如果有就可以构建均值回归的头寸，以期望在他们恢复到预期的相对定价过程中套利。

下一章中我们将概要介绍一个配对交易策略的理论过程模型，大致框架如下:

- 发掘相关系数高的资产配对
- 选择潜在的盈利能力选择配对
- 模拟这些配对的市场交易，生成策略报告

第 2 章　配对交易理论模型

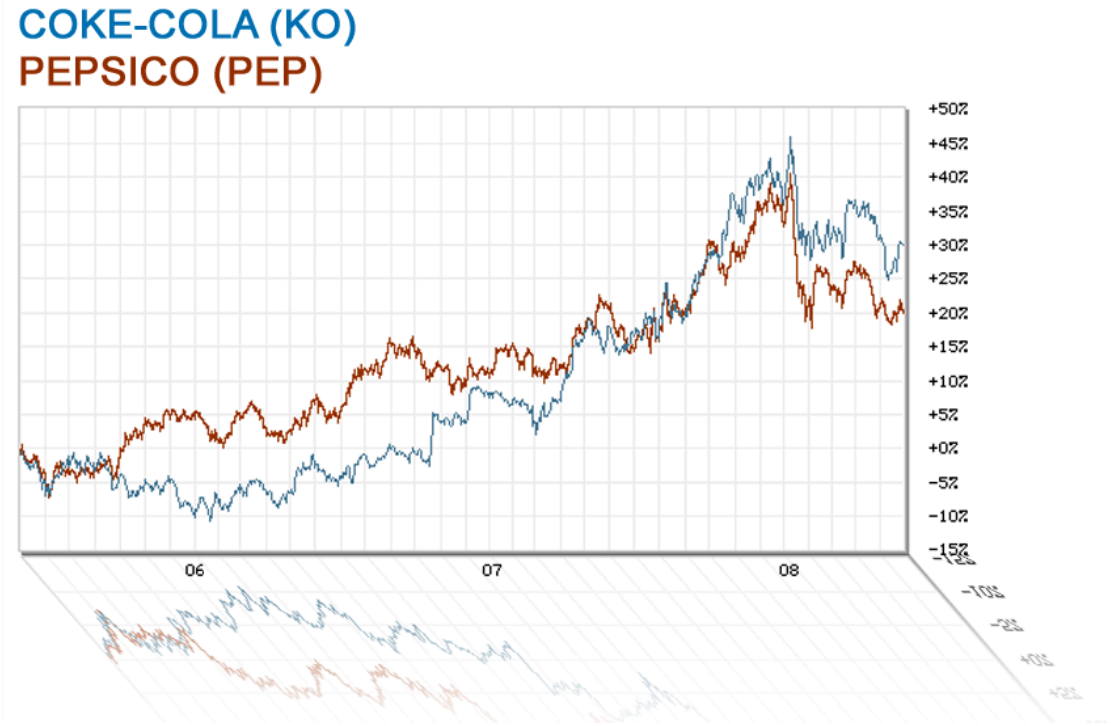

图 2.1　l 例子：高相关性的配对组合

2.1　模型假设及目的

投资的要义是购买被低估的资产和卖空被高估的资产。但是如果不知一个资产的真实价值，我们难以评估它的价格究竟是高估还是低估了它，而准确地评估一个资产的真实价值几乎是不可能的。配对交易解决这个问题的方式是采用相对定价（relative pricing）[2]，此时资产本身的名义价格将无关紧要，重要的只是两个资产的归一化（normalization）之后的价格相同。如果价格发生了分离，那么就可能是其中一个资产被高估，或者另一个资产被低估，也或者这两个事件同时发生。[3]

配对交易隐含假设了错误定价（mis-pricing）[4] 会在未来自动纠正。错误定价的意思就是上文中归一化后的价格发生分离的意思。错误定价的发生可以通过数学回归方法中的残差传布（residual spread）进行捕捉。当线性回归产生一个 beta

以及超过相对于市场的最小发散值，那么就可以开始构造一对长短头寸。如此一来，该长短头寸投资组合的回报率与市场走势无关，这就是真正意义上的市场中性策略。

不同于纯粹的经验方法，我们用经验模型和数据验证的方法定义理论估值的概念。我们将在后面表明，理论估值方法帮助我们轻松识别基于资产历史价格的配对。这也导出用于测量传布的，两个证券之间错误定价程度的公式。根据套利定价理论，如果两只股票有相同的危险因素暴露，那么预期收益是一样的。因为两个不同的证券的具体回报的实际回报可能会略有不同。让证券 A 和 B 在时间 t 的价格是 p_t^A 和 p_t^B，并且在时间 $t+i$ 分别是 p_{t+i}^A 和 p_{t+i}^B。这两个证券的回报在时间 i 会是 $log(p_t^A) - log(p_{t+i}^A)$ 和 $log(p_t^B) - log(p_{t+i}^B)$。

现在假设我们有目前时间两种证券的价格。在两个证券的回报率预计将在所有的时间框架是相同的。换言之，增量到价格的在当前时间的对数必须是在将来所有的时间实例大致相同的。这当然意味着，时间系列里的 2 只资产价格的对数必须一起移动，因此，传布计算公式是基于价格的对数之差。

解释完方法，我们现在需要在精确的术语来定义我们的意思，即两个证券的价格序列或价格对数的序列必须一起移动。两个时间序列的共同运动的理念在计量经济学领域已有很好的发展。我们在协整下一节重新讨论这个问题。

2.2 协整（相关程度）

在时间系列中，我们简要介绍非平稳序列的预处理步骤。该系列通常是由差分转化为平稳时间序列。推而广之，当分析多元时间序列，其中每个部分的系列是不固定的，这将随后区别各组分，然后逐一进行分析。

现在我们更正式地定义协整。设 y_t 和 x_t 是两个非平稳时间序列。如果存在某定值 γ 使得该系列 $y_t - \gamma x_t$ 是静止的，那么这两个序列被说成被共整合。协整的现实经济学生活中的例子比比皆是。事实上，协整的第一示范和试验涉及经济变量的配对如消费和收入，短期和长期利率，M2 货币供应量和 GDP，等等。

对于共整合的解释是由纠错捕获。这个想法是，协整系统具有长期均衡。正规定理指出纠错和协整关系是等价的陈述。我们不会试图讨论定理的证明，而只是在这里提出供参考。[1]

令 ε_{x_t} 为对应时间序列 $\{x_t\}$ 的白噪声过程。令 ε_{y_t} 为对应时间序列 $\{y_t\}$ 的白噪声

过程。纠错表示为

$$\begin{aligned} y_t - y_{t-1} &= \alpha_y(y_{t-1} - \gamma x_{t-1}) + \varepsilon_{y_t} \\ x_t - x_{t-1} &= \alpha_y(y_{t-1} - \gamma x_{t-1}) + \varepsilon_{x_t} \end{aligned} \tag{2-1}$$

解释上面的方程，左式是每个时间步的递增时间序列，右侧是纠错部分和白噪声部分的总和。看第一个方程的纠错部分 $\alpha_y(y_{t-1} - \gamma x_{t-1})$。$y_{t-1} - \gamma x_{t-1}$ 代表整个偏离长期均衡（均衡值在此情况下为零），γ 是协整系数。α_y 是纠错率，反映速度与时间序列修正自身以保持平衡。因此，这两个系列的演变随着时间的推移，从长期均衡偏离是由白噪声引起的，而这些偏差都在未来的时间步长随后纠正。

现在我们将说明，纠错的想法确实导致一个稳定的时间序列传布。生成两个独立的用零均值和单位标准差来表示的白噪声序列 ε_{y_t} 和 ε_{x_t}。其他的值设置为 $\alpha_y = -0.2, \alpha_x = 0.2$ 和 $\gamma = 1.0$。请注意，有两个系数是非常重要的 α_y 和 α_x 设置为符号相反的纠错行为。然后，利用模拟数据和方程从纠错的表示生成了两个时间序列 $\{x_t\}$ 和 $\{y_t\}$ 的值。这两个系列的情节显示如下图。接着，扩散在每个

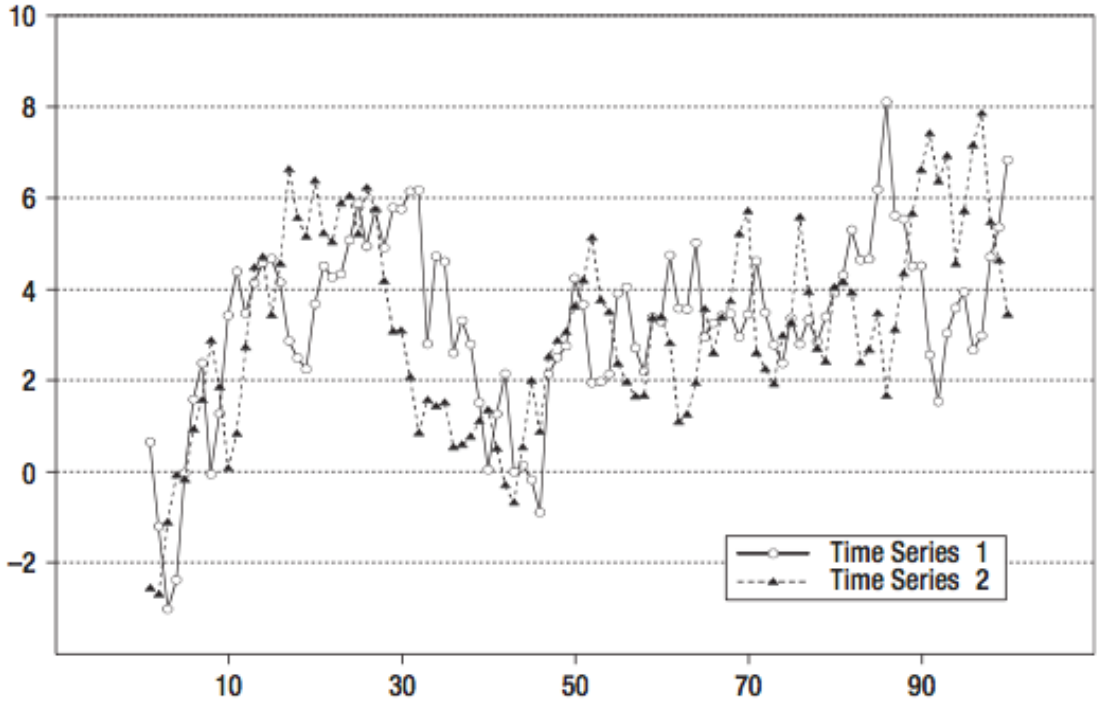

图 2.2　模拟股票价格的时间序列

时间实例使用已知值 γ 计算。传布系列和其自相关 [4] 如下图所示。 一个更直

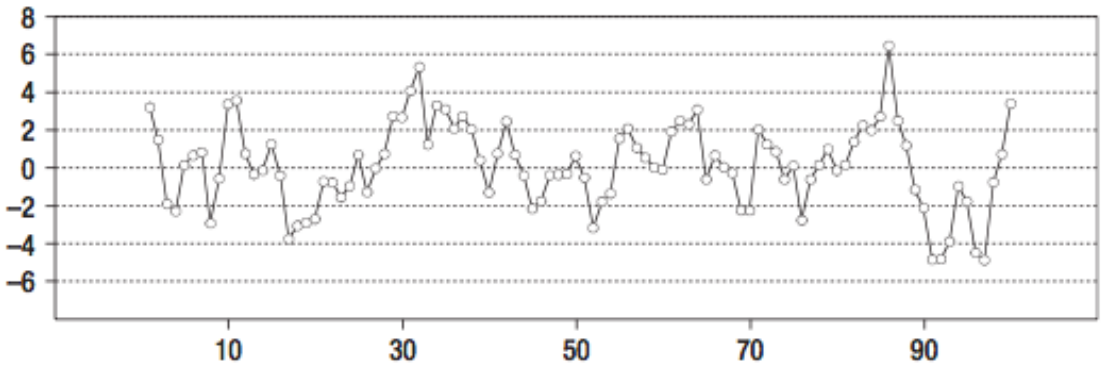

图 2.3　残差传布的时间序列

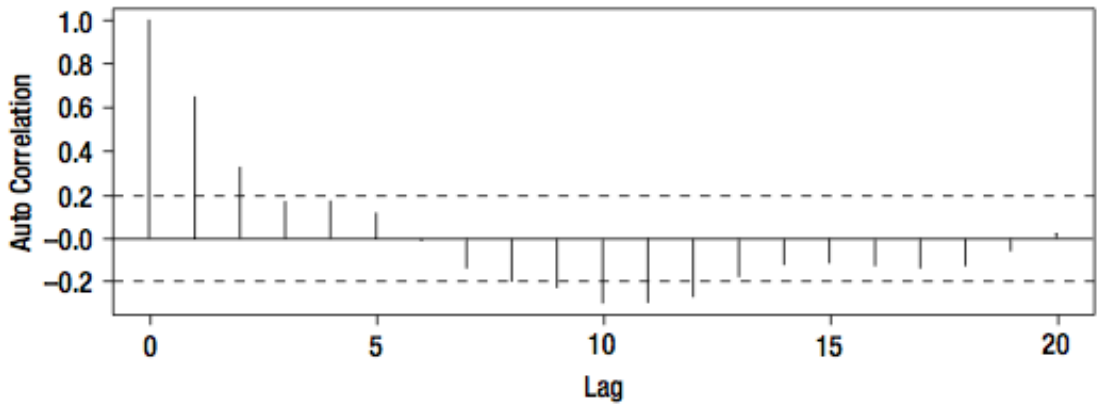

图 2.4　残差传布的自回归

接的方法来模拟股票和 Watson 的协整，是共同趋势模型 [3]。常见的趋势模型的主要思想是，一个时间序列被表示为两个分量的时间序列的简单总和：一个固定元件和一个非固定的组分。如果两个系列是协整的，那么协整线性成分的作用是抵消了非平稳成分，只留下静止部件。要明白我们的意思，考虑两个时间序列

$$\begin{aligned} y_t &= n_{y_t} + \varepsilon_{y_t} \\ z_t &= n_{z_t} + \varepsilon_{z_t} \end{aligned} \tag{2-2}$$

其中 n_{y_t} 和 n_{z_t} 是随机游走（非静止的）的两个时间序列的成分，ε_{y_t} 和 ε_{z_t} 是在协整组合。另让线性组合 $y_t - \gamma z_t$ 是共同集成的组合，导致一个固定的时间序列。扩展线性组合并整理术语，我们有

$$y_t - \gamma z_t = (n_{y_t} - \gamma n_{z_t}) + (\varepsilon_{y_t} - \gamma \varepsilon_{z_t}) \tag{2-3}$$

如果在上面的方程中的组合必须是静止的，非固定的组分必须是零，这意味着 $n_{y_t} = \gamma n_{z_t}$，或一个系列的趋势部件[5] 必须在其他系列的趋势分量的标量倍数。因此，对于两个系列进行共整合的趋势必须相同高达一个标量。我们将依靠股票 - 沃森模型 [7]，建立协整关系。

2.3 剩余利差（事件定义）

在本节中，我们用股票价格的对数适应协整模型。[1] 对于积模型参数的应用中，我们需要的股票价格的对数是一个非平稳序列。假设股票价格的对数是一个随机游走（非平稳）是标准之一。

两只股票 A 和 B 共同结合的非平稳时间序列对应 $\{log(p_t^A)\}$ 和 $\{log(p_t^B)\}$。应用纠错表示，我们有

$$\begin{aligned} log(p_t^A) - log(p_{t-1}^A) &= \alpha_A log(p_{t-1}^A) - \gamma log(p_{t-1}^B) + \varepsilon_A \\ log(p_t^B) - log(p_{t-1}^B) &= \alpha_B log(p_{t-1}^A) - \gamma log(p_{t-1}^B) + \varepsilon_B \end{aligned} \tag{2-4}$$

唯一确定模型的参数是协整合高效的 γ 和两个纠错常数 α_A 和 α_B。因此，估计该模型涉及的值确定为 α_A, α_B 和 γ。上述公式的左侧是股票在当前时间段的收益。在右边是长期均衡 $log(p_{t-1}^A) - \gamma log(p_{t-1}^B)$。换句话说，它是价格的对数标度的差异。这跟我们前面讨论的蔓延是一样的。下标为股票价格的长期均衡表达式为 $t-1$。从平衡过去的偏差决定在时间序列中的下一个点的作用。因此，在过去的变现的知识可被用于让我们在预测增量到价格的对数的边缘，也就是回报。

考虑一个投资组合，该组合买入一股 A 和卖空 γ 股 B。该投资组合在固定时间的回报率是

$$[log(p_{t+i}^A) - log(p_t^A)] - [log(p_{t+i}^B) - log(p_t^B)] \tag{2-5}$$

经演算，我们发现上述式子等同于

$$spread_{t+i} - spread_t \tag{2-6}$$

因此，投资组合的回报率是第 i 个投资时间段传布值的增量。我们已成功的将一个投资组合和一个平稳的时间序列联系在一起。接下来的事情就是提供一个解释给协整系数 γ。

2.4 交易例子

现在我们建立一个简单的交易策略例子。这个想法是基于传布的均衡值的震荡交易。我们可以根据偏差换上了从贸易平衡值和放松的贸易平衡时恢复。因此，考虑到传播的波动同样在左右平衡值的两个方向，我们有可能放松的贸易时的传播偏离在其他方向。这是由两个因素降低平均交易频率。鉴于个股有买卖差价，我们会在每次执行一个交易时间招致贸易打滑，降低了交易次数减少这种滑动的效果。

我们考虑一个策略，该策略的展开与关闭基于偏离长期均衡 μ 一个 Δ 的距离。当时间序列低于期望值 Δ 的时候，我们买入投资组合（做多 A 和卖空 B)。反之当时间序列高于期望值 Δ 时，我们卖空投资组合（走多 B 和卖空 A)。[6]

$$\begin{aligned} log(p_t^A) - \gamma log(p_t^B) &= \mu - \Delta \\ log(p_{t+i}^A) - \gamma log(p_{t_i}^B) &= \mu + \Delta \end{aligned} \tag{2-7}$$

该交易的收益传布的变化量，也就是 2Δ。

假设两只协整的股票 A 和 B 有以下参数:

$$\begin{aligned} \text{Cointegration Ratio} &= 1.5 \\ \text{Delta used for trade signal} &= 0.045 \\ \text{Bid price of A at time } t &= \$19.50 \\ \text{Ask price of B at time } t &= \$7.46 \\ \text{Ask price of A at time } t+i &= \$20.10 \\ \text{Bid price of B at time } t+i &= \$7.17 \\ \text{Average bid ask spread for A} &= 5 \text{ bps} \\ \text{Average bid ask spread for B} &= 10 \text{ bps} \end{aligned} \tag{2-8}$$

我们首先根据买卖价差检验该策略是否可行。

$$\text{Average trading slippage} = (0.0005 + 1.5 \times 0.0010) = 20 \text{ bps}$$

这比 delta 值 0.045 小，因此交易可行。

在时间 t，买入 A 股并卖空 B 股的比例为 1 : 1.5，则

$$\text{Spread at time } t = log(19.50)1.5 \times log(7.46) = 0.045$$

在时间 $t+i$，卖出 A 股同时买回 B 股，则

$$\begin{aligned} \text{Spread at time } t &= log(20.10)1.5 \times log(7.17) = 0.045 \\ \text{Total return} &= \text{return on } A + \gamma \times \text{return on } B = 0.09 \end{aligned} \tag{2-9}$$

第 3 章　实现配对交易模型的算法

目前为止，我们概括了一旦发现两协整的股票是如何进行他们的交易以实现配对交易的目的。但前面的概括也带出了很多细节方面的问题。例如我们如何定义股票配对，怎么样才能算协整的股票对？如何定义协整系数？决定打开和关闭头寸的 Delta 值应设为什么值比较合适？我们在这一章中将分小节探讨这些问题。在该章的结束，我们会展现设计的一条路径以及配对交易策略的分析。

这些步骤包括:

- 识别潜在协整的股票配对。这个过程基于股票的基础属性，即用纯统计的方法利用历史数据来判断。我们假设了过去高度协整的股票对在未来同样保证相同的属性，该性质已用时间序列的自回归检验。
- 一旦潜在的配对被识别了，我们就可以检验上述的假设，即过去高度协整的股票配对在未来同样保持高度协整的性质。这需要决定协整系数和生成传布的时间序列，才能确认它是否是平稳和均值回归的。
- 在这之后，我们可以通过分析高协整的配对来决定 delta 值。一个合理的 delta 应该略大于买卖价差导致的滑移，在决定 delta 的定义后也会给出持有期的定义，也即决定什么时候关闭头寸。

3.1　寻找并选择配对

配对交易的第一步是发现高相关的股票。在我的实现过程中是使用过去一段时间历史价格的相关系数进行定义，这一过程会在一股票集中的任两只股票遍历。在此论文中股票集以 S&P 100 为例子。一旦计算出所有股票对的相关系数矩阵，我们可以用一限定值来挑选那些高相关系数的配对（约为 0.97)。下表包含了 10 对最高相关性的配对。

有了这一系列高相关的配对，我们可以计算他们在过去固定时间段内的总回报率，然后汇出每个配对的回报图，并且找到最适合他们数据的回归线（见下图)。这条最适回归线就是我们将用未来的回报率来进行判断离散大小的基线。在我们限定的时间范围内，我们会观察每一天相对于起初的总回报率，并且计

表 3.1　10 对最高相关性的配对

Index	Equity 1	Equity 2	Correlation
1	LMT	RTN	0.994625
2	LMT	SBUX	0.982215
3	GD	LMT	0.982196
4	MET	WFC	0.979455
5	MMM	TXN	0.978430
6	FOXA	TWX	0.978277
7	BA	SBYX	0.978092
8	GD	RTN	0.978003
9	TXN	UTX	0.977723
10	HON	MMM	0.977600

算出他们相对于基线的距离（残差）的标准差，该值将会被用作决定打开头寸的信号。而基线也将用来决定什么时候关闭头寸。

有了基线和配对回报率每一天的点相对基线的距离的标准差（该距离称为残

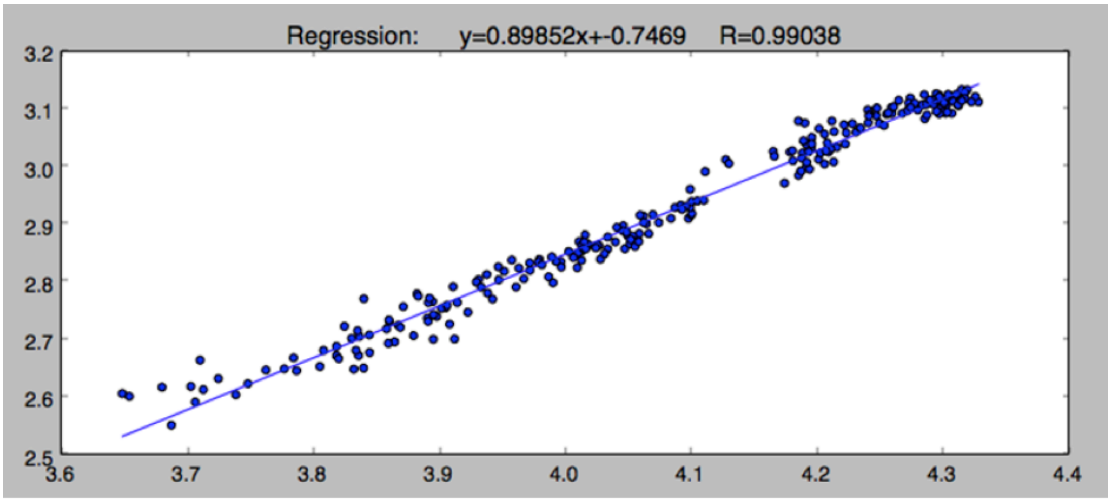

图 3.1　一对高相关性配对的基线 (APD 和 CSCO)

差)，我们可以在残差大于一个标准差值的时候打开头寸，期望其回归均值的过程能帮助我们挣得配对的回报。当残差的时间序列回复到 0 时我们就关闭头寸。下图是一个配对的残差时间序列的例子。

为了决定哪些配对是最高收益的，注意配对的回报率基于亮点：

- 残差的标准差相对于基线的距离

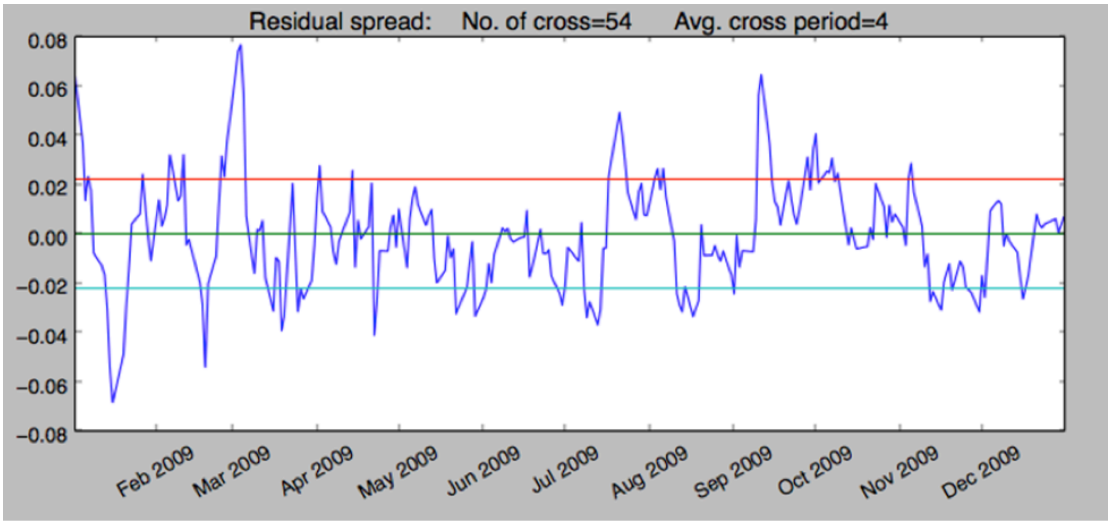

图 3.2　一对高相关性的配对的残差传布，其中其一个标准差的线也标明

- 残差传布的时间序列跨过基线的频率

二者的乘积反应了一组配对的收益能力，因此我们希望找到那些在历史上就体现出高收益能力的配对，并将在未来的交易中继续使用这些配对。下表是 10 对最具收益能力的配对。

表 3.2　list of top 10 most profitable pairs on Dec 10th, 2013

Index	Equity 1	Equity 2	Correlation	Profitablity
1	DD	GILD	0.971380	69.62
2	MDT	MET	0.971099	63.47
3	FOXA	GILD	0.971349	61.75
4	GD	RTN	0.978003	56.64
5	LMT	SBUX	0.982215	49.76
6	FOXA	TWX	0.978277	46.04
7	RTN	SBUX	0.972998	39.98
8	GD	LMT	0.982196	39.90
9	COP	MA	0.974329	36.99
10	TXN	UTX	0.977723	32.91

3.2　后台测试

下面是基于前面小节的交易策略实现：

- 每个交易日，我们遍历 S&P 100 中每两只股票组成的配对，并选择其中收益能力最好的高相关配对。然后检验他们之中是否有一对的相对价格残差穿过了基线一个标准差的值，如果有我们就打开头寸。也即，如果大于 1 倍的 sigma 值就买入 A 卖空 B，如果低于 1 倍的 sigma 值就买入 B 卖空 A。
- 当残差的时间序列穿过基线，或者持有期超过一个月的时候，我们关闭头寸。
- 为了更加真实的模拟现实交易，我们在实现中也会考虑交易成本模拟佣金，滑移罚金和卖空利率。按照实际交易标准，佣金设为每股 $0。0035，最低限度为每次交易 $2.95, 滑移设为每次交易 5 bps(实际交易价格会在打开和关闭头寸时造成 0.05 的亏损)，卖空年利率设为 3.5%。

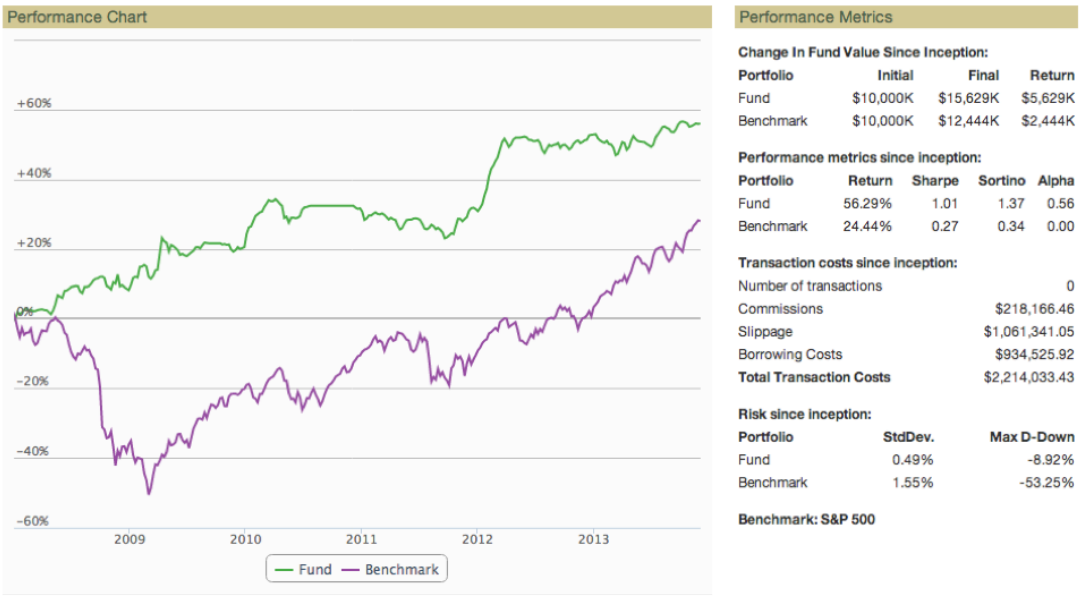

图 3.3　Performance of the ODR pairs trading strategy for 2008 to 2013 with most profitable pairs.

上图为产生策略的回报，通过股票组合评估库可发现其夏普指数每年都维持在 1.0 以上，这在金融策略中是一个十分好的结果。

3.3　参数检验和对比

在我们的第二个后台测试中，我们省略一个步骤其它不变，该省略步骤就是找出最高收益能力的配对，我们就单纯用那些最高相关性的配对，观察其表

现如何。如下图所示，后台测试的结果仍然很不错但是相对于考虑收益能力参数的上图回报有所降低，说明考虑收益能力该参数是有意义的。

最后为了进一步对比证明相关系数指数也是有意义的，我们再跑一个后台测

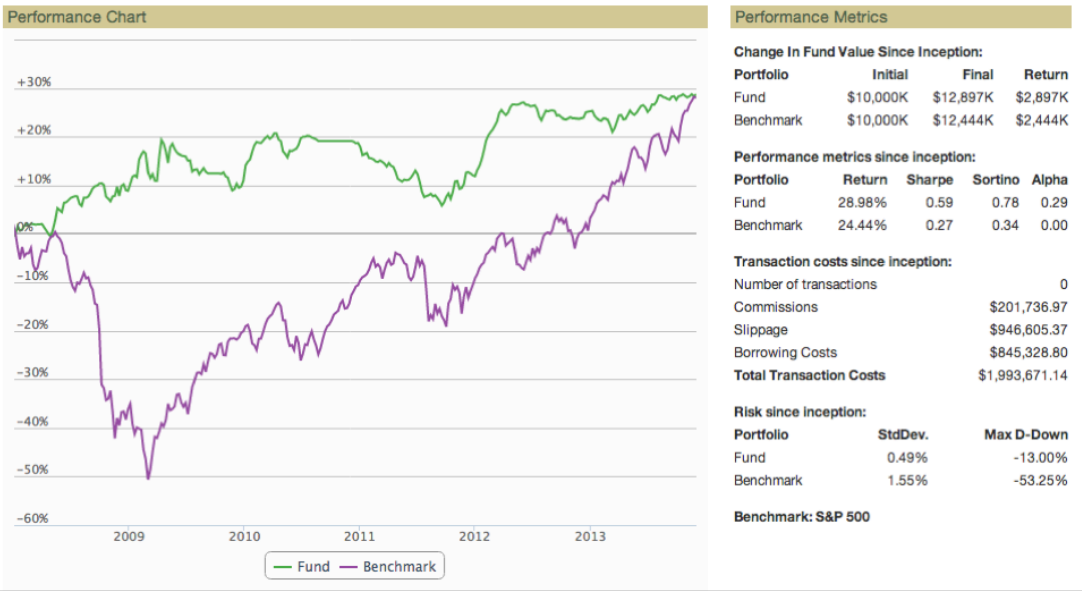

图 3.4　Performance of the ODR pairs trading strategy with most correlated pairs.

试，该后台测试只随机选择高相关的配对进行交易。该后台测试的结果如下图：

可以发现，随机选择的配对的表现远远差于前面的测试例子。关于每一天的交易情况和对每年的细致分析，将在附件中以报告形式展示。

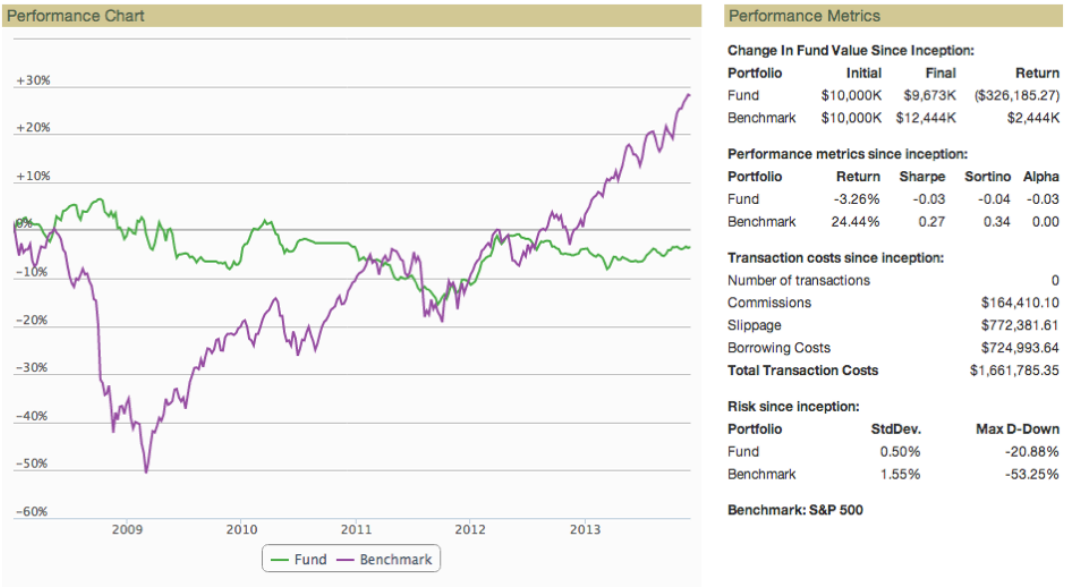

图 3.5　Performance of the ODR pairs trading strategy in which correlated pairs were selected randomly.

第 4 章　讨论和下一步研究

4.1　讨论

上述篇章我们提出了一个配对选择和基于相对价格的事件定义标准，假设错误定价会在未来进行自我纠正。该事件预测那些最近表现差的股票相对于它同属性的股票将会表现得更好。

该标准在配对交易策略中被使用，因此我们打开同等比例的长头寸和短头寸配对中的股票。该策略进行了三种不同的后台测试，分别使用——

- 最高收益能力的配对；
- 最高相关性的配对；
- 随机相关性的配对；

当我们一步一步的移去挑选配对的收益能力（第二个）和相关性排序（第三个），后台测试相对于最开始的完成策略明显的一个比一个表现得差。这说明我们采纳用来挑选配对的信息都是有意义的。

4.2　进一步研究

其实还有一些重要的因素在验证过程可以研究，例如是时间分段、市场趋势等。关于时间分段，如果我们专门只看一组配对，并且记录它是如何成功的收敛的。通过这个研究，我们将可以更成功的完善配对逃选标准以进一步选择真正更高收益能力的。关于市场行为，由我们现在的研究可以看出，配对交易策略在市场是熊市的时候表现得十分优异，原因很简单因为它是一个市场中性策略，但另一方面 2013 年的市场是个大牛市，配对交易策略想赶上市场的高回报是很难的。

另外，还有一些其它重要的因素可以进一步优化，下面不做阐述，仅简单列举——

- 股票集 (S&P 500 vs. S&P 100：更大的股票集当然更好，因为能选出更优质的配对，但会大大增加计算机的压力，因为相关系数矩阵的计算复杂度为 $O(N^2)$；另外还需要考虑每只股票卖空的限制，S&P 100 一般都能以低卖空利率进行交易，而 S&P 500 不一定)

- 回顾时间段: 我们构造基线和计算相关系数时使用的时间是 1 年，有必要对它的可靠性进行分析
- 引入排名时间段：更可靠的寻找高收益能力的配对应当适当选择另一个时间对配对交易的回报进行实际计算
- 相关系数（协整系数）的阈值：不同阈值对收益的敏感程度

第 5 章　Python 代码库的实现

为了进行策略的分析，需要构建一些金融分析库，本章就写过的 Python 库作介绍。该研究基于计算机模型和语言进行金融分析和预测，算法上主要采用 Machine Learning 的方法进行股票预测。下面分七小章进行介绍，本人担心如不是兼具金融和计算机知识的人，看内容可能会比较晦涩，所以把通俗易懂的代码库功能写在代码库具体实现方法之前，如只需了解函数的输入与输出，可跳过实现方法。

5.1　建立 QSTK 开发环境

QSToolKit 是基于 Python 的开源构架，设计目的是为了支持投资组合的建立和管理。导师初步建立开发 QSToolKit 主要是供金融、计算机的学生和有编程经验的数量分析员使用。它目前还不是一个桌面开发平台，仅仅只是是应用的基础架构，支持模型、测试和交易的工作流程。代码层面的第一个任务就是安装这套开源架构，熟悉其中的代码，为后面基于这份架构的开发和实验做准备。另注明，虽然很多代码都是运用 QSTK 中的库，但本人也是 QSTK 的开发者之一。

5.2　投资组合的评估和优化

5.2.1　评估优化库的调用方法

- 学习用 QSTK 代码进行股票价格数据的时间序列分析。我们使用的编程语言是 Python，它既不会象 C 那么复杂，同时几行代码就能实现十分强大的功能，而且本身就有很多库支持数据的操作和展示。代码会作为附件一同提交，但在报告中论文代码没太大意义，这里用几幅图说明这部分的工作。

  以上都是用 Python 的 Plot 库实现的图案绘制，上图是各股票基于首日价格的价格走线图，下图是日回报图。

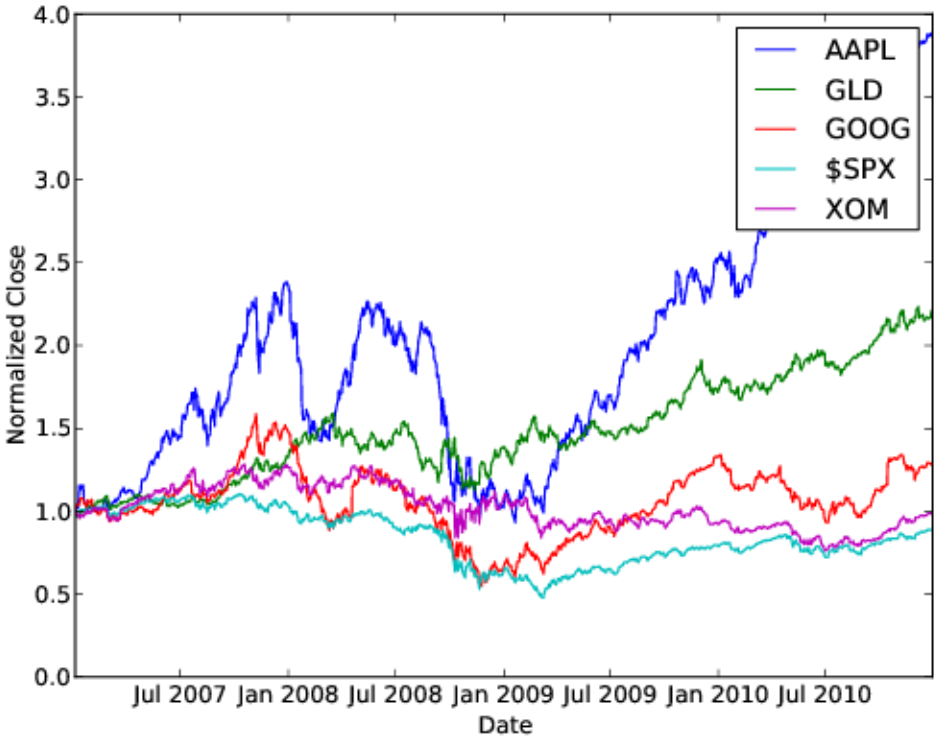

图 5.1　Normalized Close

- 写一个 Python 函数，它可以模拟包含 4 支股票的投资组合过程和评估该投资组合的表现。函数的输入是开始时间、结束时间、股票代码（如 GOOG、AAPL、GLD、XOM）和在模拟投资开始时在各股票上的资产分配比例（如 0.2、0.3、0.4、0.1），函数的输出是投资组合日回报的标准差、平均日回报、夏普比例和总回报。举个例子，调用函数时可以输入代码

```
vol, daily_ret, sharpe, cum_ret = simulate(startdate,
   enddate, ['GOOG','AAPL','GLD','XOM'],
   [0.2,0.3,0.4,0.1])
```

- 用上述的 similate() 函数做一个投资组合优化器。简单地说，就是用一个 for 循环把所有在 4 支股票上“合法”的资产分配集合元素都尝试一遍，纪录“最优”的投资组合，并作为输出打印出来。

  “合法”分配的意思是分配比例之和为 1，并且分度值是 10%，如 $[1.0,0.0,0.0,0.0],[0.1,0.1,0.1,0.7]$。

  “最优”投资组合的意思是使得夏普比例（Sharpie Ratio）最高的投资组合。

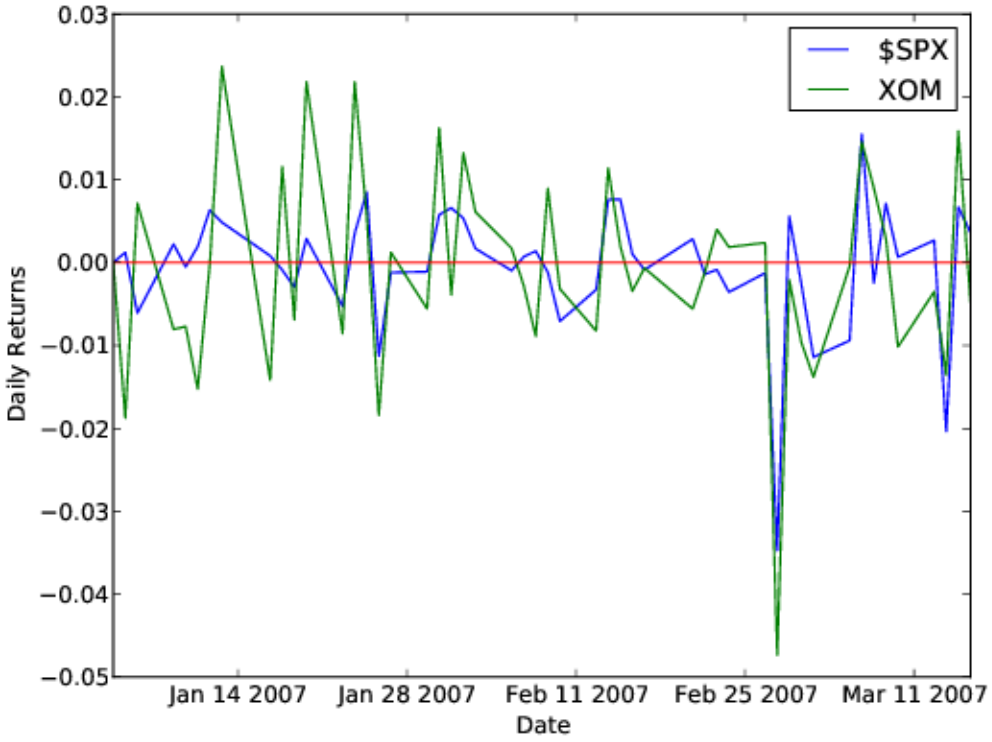

图 5.2　Daily returns

- 绘制最佳投资组合的图像，并将其和 SPY 指数作对比。

5.2.2　评估优化库的实现

了解股票历史数据的获得方法，以及如何使用 Python 和它的库 Numpy 进行股票组合的优化。QSTK 的输入

```
import QSTK.qstkutil.qsdateutil as du
import QSTK.qstkutil.tsutil as tsu
import QSTK.qstkutil.DataAccess as da
```

第三方库的输入

```
import datetime as dt
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
```

模拟函数中包括预测投资组合的回报率，并且评估其标准差、平均日收益和夏普指数。

```
def simulate(na_rets, lf_alloc):
   ''' Simulate Function'''

   # Estimate portfolio returns
   na_portrets = np.sum(na_rets * lf_alloc, axis=1)
   cum_ret = na_portrets[-1]
   tsu.returnize0(na_portrets)

   # Statistics to calculate
   stddev = np.std(na_portrets)
   daily_ret = np.mean(na_portrets)
   sharpe = (np.sqrt(252) * daily_ret) / stddev

   # Return all the variables
   return stddev, daily_ret, sharpe, cum_ret
```

主函数，包括如何定义证券集、时间序列、获取时间价格、读数据存于数据结构、模拟函数的调用和绘图，以下分块进行解释。
定义函数集：

```
   ls_symbols = ['AAPL', 'GOOG', 'IBM', 'MSFT']
```

定义时间序列的开始与结束：

```
   dt_start = dt.datetime(2010, 1, 1)
   dt_end = dt.datetime(2010, 12, 31)
```

我们使用收盘价进行分析因此每天的时间设为 16 点：

```
   dt_timeofday = dt.timedelta(hours=16)
```

调用 QSTK 中的函数获取开始和结束日期之间所有的时间节点：

```
   ldt_timestamps = du.getNYSEdays(dt_start, dt_end,
      dt_timeofday)
```

从 Yahoo Finance 获取股票数据：

```
   c_dataobj = da.DataAccess('Yahoo')
```

为了读取数据，我们还需要定义一些 Keys 以决定读取的是开盘、收盘抑或其它数据：

```
ls_keys = ['open', 'high', 'low', 'close', 'volume',
   'actual_close']
```

正式读取数据，并且存在 d_data 之中：

```
ldf_data = c_dataobj.get_data(ldt_timestamps, ls_symbols,
   ls_keys)
d_data = dict(zip(ls_keys, ldf_data))
```

拷贝收盘数据：

```
df_rets = d_data['close'].copy()
```

由于数据中可能存在缺失，因此用前一天的数据进行填补，如果不填补后面计算讲可能出现 NaN：

```
df_rets = df_rets.fillna(method='ffill')
df_rets = df_rets.fillna(method='bfill')
```

调用 simulate() 函数进行投资组合的评估：

```
na_rets = df_rets.values
na_rets = na_rets / na_rets[0, :]

lf_alloc = [0.0, 0.0, 0.0, 0.0]
max_sharpe = -1000
final_stddev = -1000
final_daily_ret = -1000
final_cum_ret = -1000
best_portfolio = lf_alloc

for i in range(0, 101, 10):
   left_after_i = 101 - i
   for j in range(0, left_after_i, 10):
      left_after_j = 101 - i - j
      for k in range(0, left_after_j, 10):
         left_after_k = 100 - i - j - k
         lf_alloc = [i, j, k, left_after_k]
         lf_alloc = [x * 0.01 for x in lf_alloc]
         stddev, daily_ret, sharpe, cum_ret =
            simulate(na_rets, lf_alloc)
         if sharpe > max_sharpe:
```

```
                max_sharpe = sharpe
                final_stddev = stddev
                final_cum_ret = cum_ret
                final_daily_ret = daily_ret
                best_portfolio = lf_alloc
```

并输出数据:

```
print "Symbols : ", ls_symbols
print "Best Portfolio : ", best_portfolio
print "Statistics : Std. Deviation : ", final_stddev
print "Statistics : Daily Returns : ", final_daily_ret
print "Statistics : Cum. Returns : ", final_cum_ret
print "Statistics : Sharpe Ratio : ", max_sharpe
```

评估投资组合的总回报率:

```
na_portrets = np.sum(na_rets * best_portfolio, axis=1)
na_port_total = np.cumprod(na_portrets + 1)

na_market = d_data['close']['SPY'].values
na_market = na_market/na_market[0]
```

图像绘制和存储:

```
plt.clf()
plt.plot(ldt_timestamps, na_port_total, label='Portfolio')
plt.plot(ldt_timestamps, na_market, label='SPY')
plt.legend()
plt.ylabel('Returns')
plt.xlabel('Date')
plt.savefig('sec1.pdf', format='pdf')
```

5.3 事件分析库

5.3.1 事件分析库的调用方法

- 事件分析库的代码在 QSTK 中已有部分未完成的雏形，经过加工和完善后，它能自主描述市场事件，然后从统计的角度观察这些事件如何影响股票价格。事件分析器采用的算法是扫描一遍特殊事件的历史数据，然后计算该事件在股票价格过去和基于一段回顾的未来的影响。
- 做一个在 S&P500 指数的专有“已知”事件的事件分析，再比较它对两组

不同领域股票的影响。事件的定义是股票的实际收盘价跌下 $5。严格地说，如果 price[t - 1] ≥ 5, price[t] < 5，那么该事件发生于 t 时刻。将这个事件分析在时间区间 2008.1.1 到 2009.12.31 测试，对比两种 S&P500 清单的结果：

- S&P500 采用 2008 年 500 支股票的清单
- S&P500 采用 2010 年的 500 支股票的清单
- S&P500 是美国股票的一个指数，综合了股市中 500 支股票的信息，并且每年这 500 支股票都有小许变化，因此 08 和 10 年会是不同的股票清单。

时间分析的结果很显然会因为采用清单的不同而有所变化，经讨论可能是以下原因：

- 我们又自己定义了一些事件，并且用之前完善的事件分析器进行实验。关于这部分产生了很多自己的猜测，并和导师讨论了不少问题，由于未经验证此处不作具体内容的阐述，但就我笔记上和导师讨论过的问题作下介绍
  - 有没可能通过自己的事件分析赚钱？
  - 如果可能的话，我调研乐目前对冲基金公司采用的投资策略，讨论入市和出市的时机以及持有期的细节问题。
  - 对配对交易策略进行了抽象的风险评估。
  - 从期望的角度说，每次交易的回报会是多少？
  - 每年会有多少次机会发生定义事件？
  - 有没有什么方法可以在此基础上降低风险？
- 就之前写的程序，即一些数据的纪录，如

```
For the $5.0 event with S&P500 in 2012, we find 176
   events. Date Range = 1st Jan 2008 to 31st Dec 2009.
For the $5.0 event with S&P500 in 2008, we find 326
   events. Date Range = 1st Jan 2008 to 31st Dec 2009.
```

5.3.2 事件分析库的实现方法

这部分可基于历史信息进行事件分析，处理如何读入和处理不同类型的历史数据、如何评估事件分析结果的问题。该库可用“事件分析”评估股票信息对未来价格的影响，其中写了 QSTK 中的事件分析器的代码也将附于附件中。

该库的输入是存于 list 中带有日期的股票符号，输出是 pandas 库中的 Datamatrix 包含事件矩阵，其格式如下：

```
   |IBM |GOOG|XOM |MSFT| GS | JP |
(d1)|nan |nan | 1  |nan |nan | 1  |
(d2)|nan | 1  |nan |nan |nan |nan |
(d3)| 1  |nan | 1  |nan | 1  |nan |
(d4)|nan | 1  |nan | 1  |nan |nan |
```

定义寻找事件的函数：

```
def find_events(ls_symbols, d_data):
   ''' Finding the event dataframe '''
   df_close = d_data['actual_close']
   ts_market = df_close['SPY']

   print "Finding Events"

   # Creating an empty dataframe
   df_events = copy.deepcopy(df_close)
   df_events = df_events * np.NAN

   # Time stamps for the event range
   ldt_timestamps = df_close.index

   for s_sym in ls_symbols:
      for i in range(1, len(ldt_timestamps)):
         # Calculating the returns for this timestamp
         f_symprice_today =
            df_close[s_sym].ix[ldt_timestamps[i]]
         f_symprice_yest = df_close[s_sym].ix[ldt_timestamps[i
            - 1]]
         f_marketprice_today = ts_market.ix[ldt_timestamps[i]]
         f_marketprice_yest = ts_market.ix[ldt_timestamps[i -
            1]]
         f_cutoff = 5.0
         if f_symprice_today < f_cutoff and f_symprice_yest >=
            f_cutoff:
            df_events[s_sym].ix[ldt_timestamps[i]] = 1

   return df_events
```

调用函数得到所需的事件矩阵：

```
df_events = find_events(ls_symbols, d_data)
```

最终调用 QSTK 中的函数进行事件分析的构图：

```
    print "Creating Study"
ep.eventprofiler(df_events, d_data, i_lookback=20,
    i_lookforward=20,
            s_filename='MyEventStudy.pdf',
                b_market_neutral=True, b_errorbars=True,
            s_market_sym='SPY')
```

以下是配对交易通过事件分析库生成的图，我们可发现该策略可行:

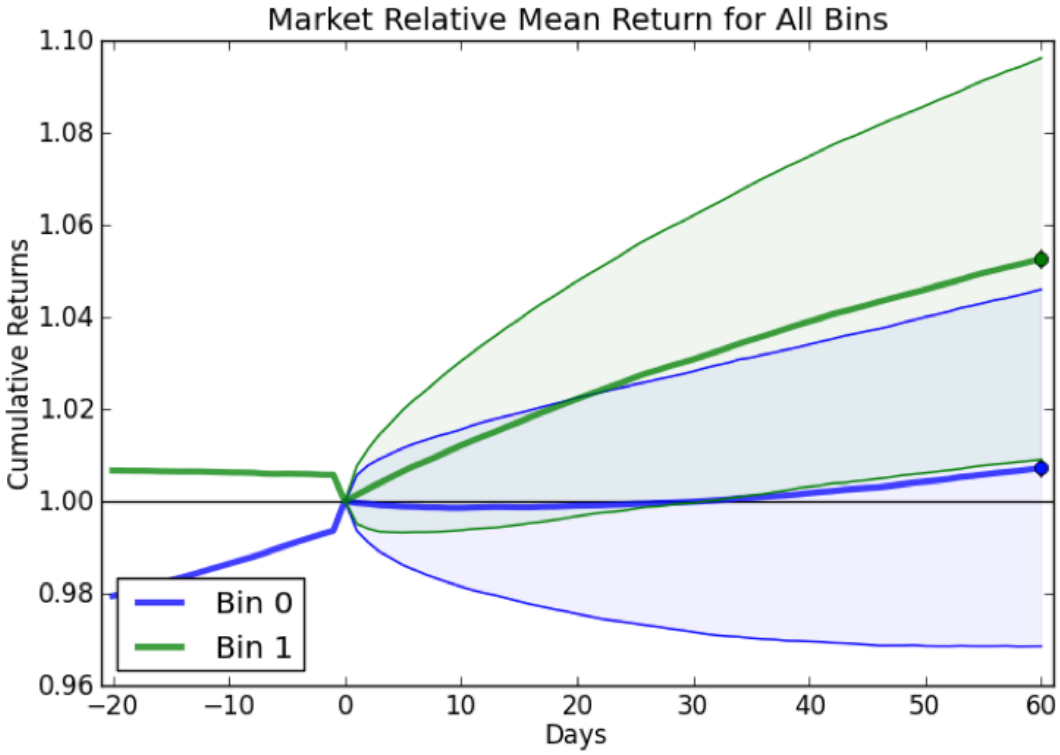

图 5.3　配对交易的事件分析

5.4　市场模拟库

5.4.1　市场模拟库的调用方法

- 制作市场模拟器，也就是写一个叫 market.py 的文件，可以接受象下面这样的一条命令行：

```
python marketsim.py 1000000 orders.csv values.csv
```

其中数字代表开始投资时刻的现金，orders.csv 是纪录了订单的输入文件，每条订单包括年、月、日、股票代码、买或卖和交易股票数量，如（2008, 12, 3, AAPL, BUY, 130），（2008, 12, 8, AAPL, SELL, 130）。模拟器需要每天通过实际收盘价计算投资组合的总价值，再将结果打印在 values.csv 文件中。values.csv 的结果输出会像是（2008, 12, 3, 1000000），（2008, 12, 4, 1000010），（2008, 12, 5, 1000250）以此类推。

- 制作一个投资分析工具，叫 analyze.py，它能接受如下的一条命令行：

```
python analyze.py values.csv \$SPX
```

该工具会从 values.csv 读入每日投资组合价值，并且绘制图像输出。它会使用命令行中的股票指数符号作为基准进行比较（在这个例子中是 $SPX）。简单地说就是 analyze.py 可以绘制全交易期间的投资组合历史价值的图像，并且输出投资组合的日回报标准差、平均日回报、夏普比例和总回报，如其中一个提供的 orders.csv 作为输入的输出结果如下：

```
The final value of the portfolio using the sample file is
   -- 2011,12,20,1133860
Details of the Performance of the portfolio :

Data Range : 2011-01-10 16:00:00 to 2011-12-20 16:00:00

Sharpe Ratio of Fund : 1.21540462111
Sharpe Ratio of $SPX : 0.0183391412227

Total Return of Fund : 1.13386
Total Return of $SPX : 0.97759401457

Standard Deviation of Fund : 0.00717514512699
Standard Deviation of $SPX : 0.0149090969828

Average Daily Return of Fund : 0.000549352749569
Average Daily Return of $SPX : 1.72238432443e-05
```

以上所阐述的输入输出文件，以及 Python 代码文件将作为附件提交。

5.4.2 市场模拟库的实现方法

市场模拟器可以接受交易订单，同时记录投资组合的价值，并将之储存在.csv 文件中。同时也进行股票投资组合表现的评估。

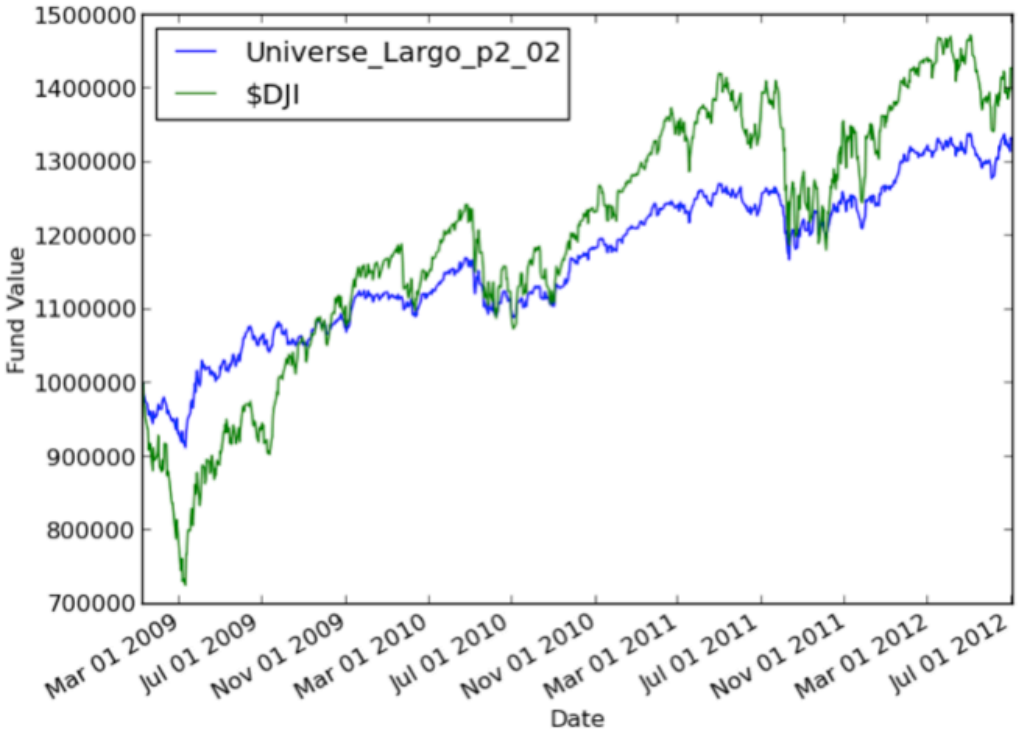

图 5.4　Fund value

analyze.py 的实现方法与第二章中投资组合的评估类似，无需再失陷，我们进行 marketsim.py 的阐述，首先读入 csv 文件要求的买卖订单：

```
def _csv_read_sym_dates(filename):
    reader = csv.reader(open(filename, 'rU'), delimiter=',')
    symbols = []
    dates = []
    for row in reader:
        if not(row[3] in symbols):
            symbols.append(row[3])
        date = dt.datetime(int(row[0]), int(row[1]), int(row[2]))
        if not(date in dates):
            dates.append(date)
    dates = sorted(dates)
    return symbols, dates
```

根据要求买卖的股票进行这些股票的数据读取：

```
def _read_data(symbols, dates):
    timeofday = dt.timedelta(hours=16)
```

```
    timestamps = du.getNYSEdays(dates[0], dates[-1] +
        dt.timedelta(days=1), timeofday)

    dataobj = da.DataAccess('Yahoo')
    close = dataobj.get_data(timestamps, symbols, "close",
        verbose=True)
    close = close.fillna(method='ffill')
    close = close.fillna(method='bfill')
    return close, timestamps
```

生成每日持有股票的函数:

```
def _share_holdings(filename, symbols, timestamps, close):
    reader = csv.reader(open(filename, 'rU'), delimiter=',')
    share_matrix = np.zeros((len(timestamps), len(symbols)))
    share_matrix = pandas.DataFrame(share_matrix,
        index=timestamps, columns=symbols)
    for row in reader:
        date = dt.datetime(int(row[0]), int(row[1]), int(row[2]))
        time_stp = close.index[close.index >= date][0]
        if row[4] == 'Buy':
            share_matrix.ix[time_stp][row[3]] += float(row[5])
        elif row[4] == 'Sell':
            share_matrix.ix[time_stp][row[3]] -= float(row[5])
    return share_matrix
```

计算每日现金数量的函数:

```
def _share_value_cash(share_matrix, close, i_start_cash):
    ts_cash = pandas.TimeSeries(0.0, close.index)
    ts_cash[0] = i_start_cash
    for row_index, row in share_matrix.iterrows():
        cash = np.dot(row.values.astype(float),
            close.ix[row_index].values)
        ts_cash[row_index] -= cash
    share_matrix['_CASH'] = ts_cash
    share_matrix = share_matrix.cumsum()
    return share_matrix
```

计算总资产金额的函数:

```
def _fund_value(share_matrix, close):
    historic = close
    historic['_CASH'] = 1
    ts_fund = pandas.TimeSeries(0.0, close.index)
```

```
    for row_index, row in share_matrix.iterrows():
        ts_fund[row_index] += np.dot(row.values.astype(float),
            close.ix[row_index].values)
    return ts_fund
```

将结果作为 csv 文件输出，以便用 analyze.py 进行分析：

```
def _write_fund(ts_fund, filename):
    writer = csv.writer(open(filename, 'wb'), delimiter=',')
    for row_index in ts_fund.index:
        row_to_enter = [str(row_index.year),
            str(row_index.month), \
                    str(row_index.day), str(ts_fund[row_index])]
        writer.writerow(row_to_enter)
```

5.5 连接事件分析库和市场模拟库

5.5.1 连接库的方法

- 重新修改事件分析器使得它能根据事件的发生输出一系列的交易订单。之前写的事件分析器只是在事件矩阵（可想象成横排是股票符号，纵列是时间的表格）中放置一个 1 标志事件的发生，现在是要让输出变成订单模式，如：

```
Date, AAPL, BUY, 100
Date + 5 days, AAPL, SELL, 100
```

- 将该输出作为输入传递给市场模拟器
- 通过市场模拟器得出交易策略的表现，如总回报、平均日回报、日回报标准差和交易时期的夏普比例。
- 根据 Tucker 的要求做了两个实验。
  - 实验一：采用在任务 2 中实现的实际收盘价 $5 事件和 2012 年的 S&P500 数据。该实验和一个博士实习生分别独立实验，Tucker 希望通过如此检查我们能否得到相同的答案以保证我们各自的事件分析器改造没有问题。当时要求的输入数据是：启动资金: $50,000；开始日期: 1 January 2008；结束日期: 31 December 2009；当事件发生的时候，在当日买入 100 股股票；持有 5 天后自动卖出。
  - 实验二：设计自己的事件和交易策略，对于两个实验都要生成图像以

便观察，同时象之前一样计算夏普比例、总回报和日收入的标准差，如下——

```
The final value of the portfolio using the sample
   file is -- 2009,12,28,54824.0
Details of the Performance of the portfolio
Data Range : 2008-01-03 16:00:00 to 2009-12-28
   16:00:00
Sharpe Ratio of Fund : 0.527865227084
Sharpe Ratio of $SPX : -0.184202673931

Total Return of Fund : 1.09648
Total Return of $SPX : 0.779305674563

Standard Deviation of Fund : 0.0060854156452
Standard Deviation of $SPX : 0.022004631521

Average Daily Return of Fund : 0.000202354576186
Average Daily Return of $SPX : -0.000255334653467
```

5.6 实现金融预测指数（机器学习特征量）

5.6.1 金融预测指数的调用方法

- 用 20 日回顾实现布林带 Indicator。编写代码生成图形展示移动平均线 (下图)，股票价格和布林带的高低线。布林带的高线代表期望值加一倍的标准差，低线代表期望值减一倍的标准差。传统上应该是 2 倍的标准差，但 Tucker 希望我把布林带做得细一点所以只用 1 倍的标准差。

- 调整输出使得 indicator 产生的值在 -1 和 1 之间。实际上，这些值是有可能被超过的，希望 +1 代表价格是高于期望一倍标准差的位置，-1 代表价格是低于期望一倍标准差的位置，为了实现这个变化，只要简单地加入下面这行代码——

```
Bollinger_val = (price - rolling_mean) / (rolling_std)
```

输出图像如下图所示。

- 实现一些我自己设计的 Indicator，并把播报值调整到 -1 和 1 之间。我进行

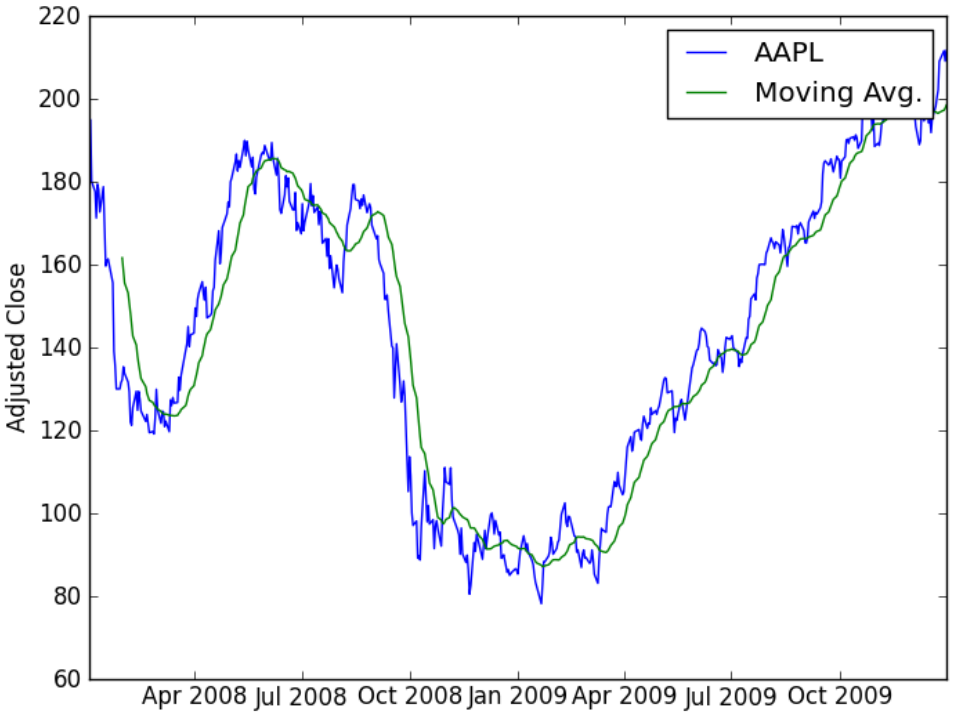

图 5.5　Bollinger band

了两个实验，实验一仍然采用上述的布林带，并同时把输出值调整到－ 1 和 1 之间。然后生成了一个时间从 Jan 1, 2010 到 Dec 31, 2010 的 GOOG 图像，即输入数据是

```
Symbol: GOOG; Startdate: 1 Jan 2010; Enddate: 31 Dec
   2010; 20 period lookback
```

实验二是关于 relative strength 的 Indicator 实现，这部分是我最近才开始的工作，也是 Lucena Research 着手使用的最新预测工具，涉及公司机密，我签署了不泄露合同所以不能在该实习报告中有所展示。但所作的工作步骤和上述布林带的实现是一样的，只是稍显复杂，以下是布林带实验的一个输出结果例子（采用 Python 中的 pandas 库可以直接计算 Bollinger band 的值)——

```
                     AAPL     GOOG     IBM      MSFT
2010-12-23 16:00:00 1.185009 1.298178 1.177220 1.237684
2010-12-27 16:00:00 1.371298 1.073603 0.590403 0.932911
```

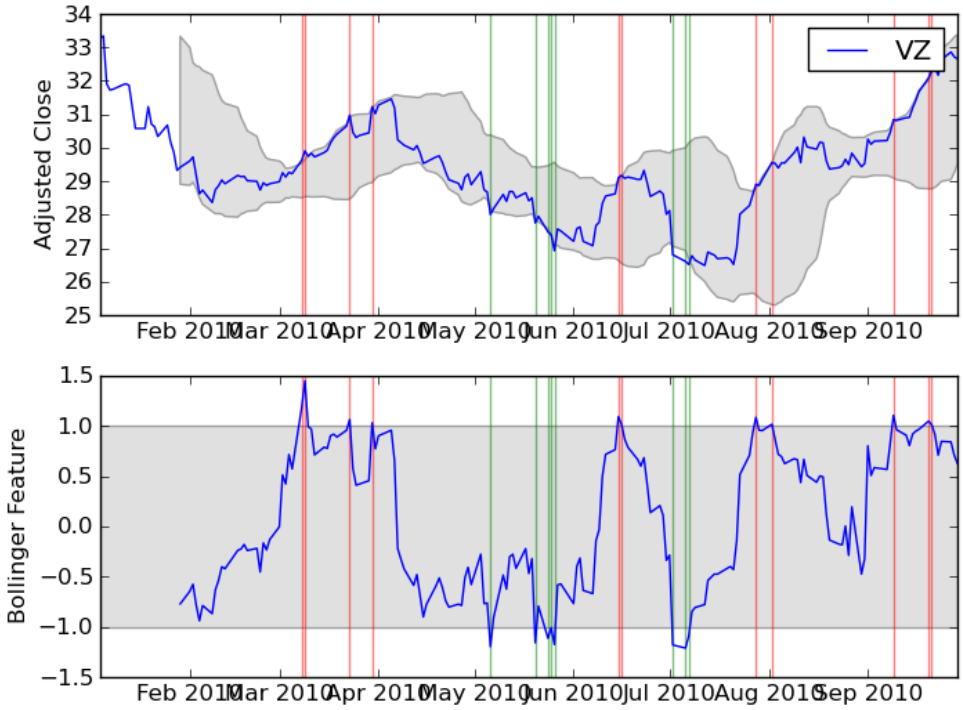

图 5.6　Normalized Bollinger band

```
2010-12-28 16:00:00 1.436278 0.745548 0.863406 0.812844
2010-12-29 16:00:00 1.464894 0.874885 2.096242 0.752602
2010-12-30 16:00:00 0.793493 0.634661 1.959324 0.498395
```

5.6.2　用布林带进行事件分析

从该部分开始 Tucker 开始让我着手 technical indicator 的研究，首先从布林带开始之后再进行一些最前卫的 indicator 实现。

- 用 20 日回顾实现布林带 indicator。还是一样的定义，布林带的高线代表期望值加一倍标准差，低线代表期望值减一倍标准差。同时把 indicator 的输出值调整在－ 1 和 1 之间。
- 采用以下特征进行事件分析：

```
Bollinger value for the equity today ≤-2.0, Bollinger
   value for the equity yesterday ≥2.0, Bollinger value
   for SPY today ≥1.0.
```

因此我们是在寻找股票跌穿布林带低线，同时市场走势却是稳定地向上走。这说明肯定有些特殊的事情发生在个股上。

- 使用上一项目做出来的 indicator，再按上述方法进行事件分析，试图挖掘一些有意思的结果。该部分的实验 Tucker 也给了统一数据如下：Event 如 2 中所写，

```
Startdate: 1 Jan 2008, Enddate: 31 Dec 2009; 20 day
   lookback for Bollinger Bands; Symbol list: SP5002012;
   Adjusted close.
```

以下是事件分析的一个结果。

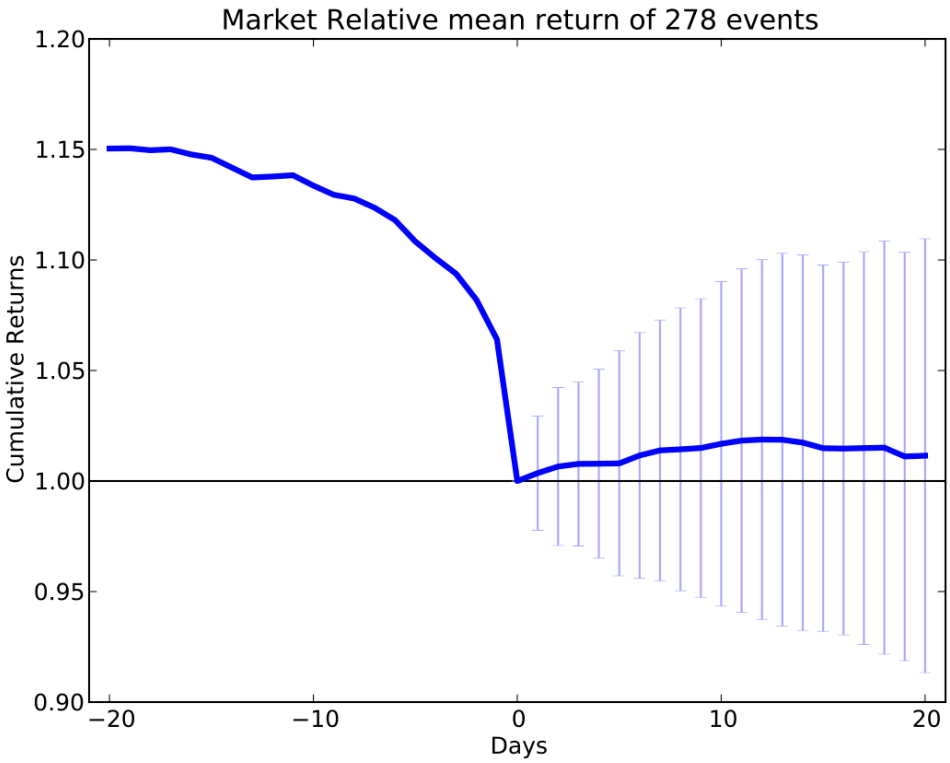

图 5.7　Event study of Bollinger band

5.7 Indicators 库

该部分代码由于太多，附在 lrlfeatures.py 中，内容和 Bollinger Band 的流程相似，只是不仅实现 Bollinger Band 这一种 indicator，而是实现了 104 种包含 384 种变化的 indicator 集合，方便以后进行事件分析的时候可以综合各种 indicators 作为机器学习特征进行预测。

插图索引

图 1.1 通过线性回归方法找到的一组配对：我们会在配对的回报率发散(价格走势相离）后打开头寸，而在回报率收敛后关闭头寸。(图片截自 Google Finance） ……3
图 2.1 l 例子：高相关性的配对组合 ……5
图 2.2 模拟股票价格的时间序列 ……7
图 2.3 残差传布的时间序列 ……8
图 2.4 残差传布的自回归 ……8
图 3.1 一对高相关性配对的基线 (APD 和 CSCO) …… 13
图 3.2 一对高相关性的配对的残差传布，其中其一个标准差的线也标明 … 14
图 3.3 Performance of the ODR pairs trading strategy for 2008 to 2013 with most profitable pairs. …… 15
图 3.4 Performance of the ODR pairs trading strategy with most correlated pairs.16
图 3.5 Performance of the ODR pairs trading strategy in which correlated pairs were selected randomly. …… 17
图 5.1 Normalized Close …… 21
图 5.2 Daily returns …… 22
图 5.3 配对交易的事件分析 …… 28
图 5.4 Fund value …… 30
图 5.5 Bollinger band …… 34
图 5.6 Normalized Bollinger band …… 35
图 5.7 Event study of Bollinger band …… 36
图 B.1 converge / diverge pair …… 48
图 B.2 Performance chart for 2008-2009 …… 49
图 B.3 2008-2013 performance …… 49
图 B.4 Knight Capital – large portion of the market …… 50
图 B.5 Performance chart for 2002-2013 …… 51

表格索引

表 3.1　10 对最高相关性的配对 13
表 3.2　list of top 10 most profitable pairs on Dec 10th, 2013..... 14

公式索引

公式 2-1 ……7
公式 2-2 ……8
公式 2-3 ……8
公式 2-4 ……9
公式 2-5 ……9
公式 2-6 ……9
公式 2-7 …… 10
公式 2-8 …… 10
公式 2-9 …… 11

参考文献

[1] Engle R F, Granger C W. Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 1987. 251–276

[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] Vidyamurthy G. Pairs Trading: quantitative methods and analysis, volume 217. New York: John Wiley & Sons, 2004

[4] Banerjee A, Dolado J J, Galbraith J W, et al. Co-integration, error correction, and the econometric analysis of non-stationary data. OUP Catalogue, 1993.

[5] Stock J H, Watson M W. Testing for common trends. Journal of the American statistical Association, 1988, 83(404):1097–1107

[6] Phillips P C, Durlauf S N. Multiple time series regression with integrated processes. The Review of Economic Studies, 1986, 53(4):473–495

致　谢

衷心感谢导师唐平中老师对本人的精心指导。他的言传身教将使我终生受益。

在美国佐治亚理工学院计算机系进行七个月的合作研究期间，承蒙 Tucker Balch 教授热心指导与帮助，不胜感激。感谢 Tucker Balch 实验室 Lucena Research 全体同学们的热情帮助和支持！

感谢 ThuThesis，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

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：陈炜亮　　日　期：2014.6.20

附录 A　外文资料阅读报告

The theme for investing is to sell overvalued and buy the undervalued equities. However, it is possible to determine that a security is overvalued only if we know the true value. But this is hard to do. Pairs trading attempts to resolve this using relative pricing. The specific price of the security will be not of importance. It is only important that the normalized prices of the two securities be the same. If the prices is different, it could be that one of the securities is overpriced, the other security is underpriced, or the mis-pricing[1] is a combination of both.

Pairs trading involves with the idea that the mis-pricing will correct itself in the future. The mutual mis-pricing between the two securities is captured by the notion of spread. A long-short position is constructed such that it has a beta and therefore minimal exposure to the market. Hence, the returns from the trade are uncorrelated to market returns, a feature typical of market neutral strategies.

A.1　Principle

Unlike the purely empirical approach, the method that we subscribe to comprises theoretical valuation concepts that are then validated with empirical models and data. We will later show that the theoretical valuation approach helps us to easily identify pairs based on the security historical price. It also leads to the formula used to measure the spread, the degree of mis-pricing between the two securities. According to arbitrage pricing theory[2], if two securities have exactly the same risk factor exposures, then the expected return is the same. The actual return may differ slightly because of different specific returns for the two securities. Let the price of securities A and B at time t be p_t^A and p_t^B, and at time $t+i$ be p_{t+i}^A and p_{t+i}^B, respectively. The return in the time period i for the two securities given as $log(p_t^A) - log(p_{t+i}^A)$ and $log(p_t^B) - log(p_{t+i}^B)$.

Now let us say that we have the prices of both securities at the current time. The return on both securities is expected to be the same in all time frames. In other words, the increment to the logarithm of the prices at the current time must be about the same for

both the securities at all time instances in the future. This, of course, means that the time series of the logarithm of the two prices must move together, and the spread calculation formula is therefore based on the difference in the logarithm of the prices.
Having explained our approach, we now need to define in precise terms what we mean when we say that the price series or the log price series of the two securities must move together. The idea of co-movement of two time series has been well developed in the field of econometrics. We discuss it in the following section on cointegration[6].

A.1.1 Cointegration (correlation)

In time series we are briefly taught the preprocessing step for non-stationary series. The series is typically transformed into a stationary time series by differencing. By extension, when analyzing multivariate time series where each of the component series is non-stationary, it would then make sense to difference each component and then subject them to examination.
Let us now state the idea of cointegration more formally. Let y_t, and x_t be two non-stationary time series. If for a certain value γ, the series $y_t \gamma x_t$ is stationary, then the two series are said to be co-integrated. Real-life examples of cointegration abound in economics. In fact, the first demonstrations and tests of cointegration involved economic variable pairs like consumption and income, short-term and long-term rates, the M2 money supply and GDP, and so forth.
The explanation for cointegration is captured by error correction[1]. The idea is that cointegrated systems have a long-run equilibrium. The formal theorem stating that error correction and cointegration are equivalent representations. We shall not attempt to discuss the proof of the theorem, but simply present here for your information.
Let ε_{x_t} be the white noise process corresponding to time series $\{x_t\}$. Let ε_{y_t} be the white noise process corresponding to the time series $\{y_t\}$. The error correction representation is

$$y_t - y_{t-1} = \alpha_y(y_{t-1} - \gamma x_{t-1}) + \varepsilon_{y_t}$$

Let us interpret the above equations. The left-hand side is the increment to the time series at each time step. The right-hand side is the sum of two expressions, the er-

ror correction part and the white noise part. Let us look at the error correction part $\alpha_y(y_{t-1} - \gamma x_{t-1})$ from the first equation. The term $y_{t-1} - \gamma x_{t-1}$ is representative of the deviation from the long-run equilibrium (equilibrium value is zero in this case), and γ is the coefficient of cointegration. α_y is the error correction rate, indicative of the speed with which the time series corrects itself to maintain equilibrium. Thus, as the two series evolve with time, deviations from the long-run equilibrium are caused by white noise, and these deviations are subsequently corrected in future time steps.

A.1.2 Residual spread (event definition)

In this section, we fit the cointegration model to the logarithm of stock prices. For the cointegration model to apply, we would require the logarithm of stock prices to be a non-stationary series. The assumption that the logarithm of stock prices is a random walk (read as non-stationary) is a rather standard one.

Let us say that two stocks A and B are co-integrated with the non-stationary time series corresponding to them being $\{log(p_t^A)\}$ and $\{log(p_t^B)\}$ respectively. Applying the error correction representation described here, we have

$$log(p_t^A) - log(p_{t-1}^A) = \alpha_A log(p_{t-1}^A) - \gamma log(p_{t-1}^B) + \varepsilon_A$$

The parameters that uniquely determine the model are the cointegration co- efficient γ and the two error correction constants α_A and α_B. Therefore, estimating the model involves determining the appropriate values for α_A, α_B, and γ. The left-hand side of the above equation is the return of the stocks in the current time period. On the right-hand side, note the expression for the long-run equilibrium, $log(p_{t-1}^A) - \gamma log(p_{t-1}^B)$, in both the equations. In words, it is the scaled difference of the logarithm of price. Incidentally, this coincides with what we termed the spread in our earlier discussion. Also notice that the subscripts for stock prices in the expression for the long-run equilibrium is $t-1$. The past deviation from equilibrium plays a role in deciding the next point in the time series. Therefore, knowledge of the past realizations may be used to give us an edge in predicting the increments to the logarithm of prices; that is, returns.

Consider a portfolio with long one share of A and short γ shares of B. The return of the

portfolio for a given time period is given as

$$[log(p^A_{t+i}) - log(p^A_t)] - [log(p^B_{t+i}) - log(p^B_t)]$$

. Rearranging the terms a little bit, we have the above equation equals to

$$spread_{t+i} - spread_t$$

Therefore, the return on the portfolio is the increment to the spread value in the time period i. We have successfully associated a portfolio with a stationary time series. The one thing that remains is providing an interpretation for γ, the cointegration coefficient.

References

[1] Donald E. Knuth. The TEXbook. Addison-Wesley, 1984.

[2] Paul W. Abrahams, Karl Berry and Kathryn A. Hargreaves. TEX for the Impatient. 1990.

[3] Gatev, Evan, G., William, N. Goetzmann, and K. Greet Rouwenhorst. Pairs Trading: Performance of a Relative Value Arbitrage Rule.1999.

附录 B　外文资料的调研

Hedge funds are notorious for keeping their cards close to their chest. The idea behind -

B.1　Pairs trading

Pairs trading is a sexy strategy: market neutral, good reason to believe pairs will return to their level of correlation with one another, works regardless of the direction of the market. Many believe it was formalized with Gerry Bamberger and Nunzio Tartaglia in the early 1980’s with Morgan Stanley’s quantitative group[1] – couldn’t find earliest publication very easily.

We implemented it (brief explanation of implementation from ppt slides explaining pairs trading) There is a brief explanation of what pairs trading is (conv / div pair picture).

图 B.1　converge / diverge pair

图 B.2　Performance chart for 2008-2009

B.2　Losing value

Look at 2008-2013 performance and saw abysmal returns starting in early 2010 relative to what we had seen in 2008-2009: what went wrong?

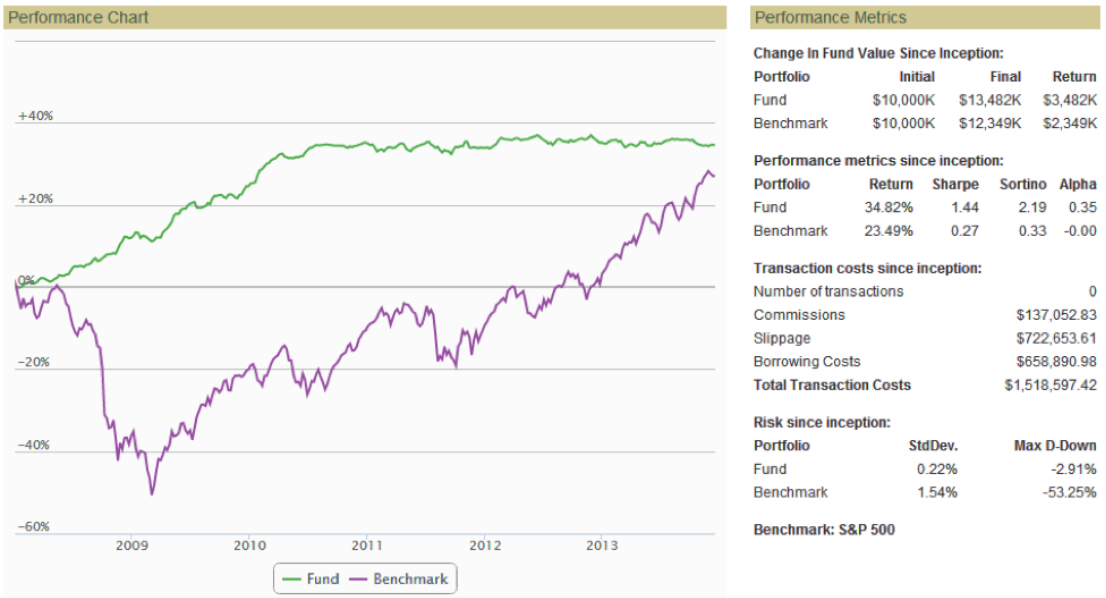

图 B.3　2008-2013 performance

B.3 Market news analysis

Looking at news events related to pairs trading, we saw a significant event in February 2010, namely, Knight Capital began using a pairs trading strategy; additionally, others began entering the picture shortly after.

- Knight Capital – large portion of the market (Figure below)
- One of the largest traders at the time with around 17% market share (both NYSE and NASDAQ[3])
- Also, Bank of America introduced pairs trading to clients later in 2010[4]
- Market share of BoA – not sure, but undoubtedly substantial
- corresponds well w/ the flattening of our performance in figures

图 B.4 Knight Capital – large portion of the market

B.4 Conclusion

It brought up the curiosity: when a book[2] was published on this topic, how did that affect the strategy’s performance? Look at 2002-2013.

It’s not a great deal of influence from the 2004 release of Vidyamurthy’s book on pairs trading. It’s also notable is the strong rally in 2009; we have theories on why this may have occurred, but we’ll save those for another time. There may be number of hedge funds dropped significantly in 2008[5]. But maybe not that much[6].

When a strategy is profitable and known, it dilutes the profits, especially as large firms enter the picture (Knight), quote from [ML guy] about market actively trying to defeat your strategy.

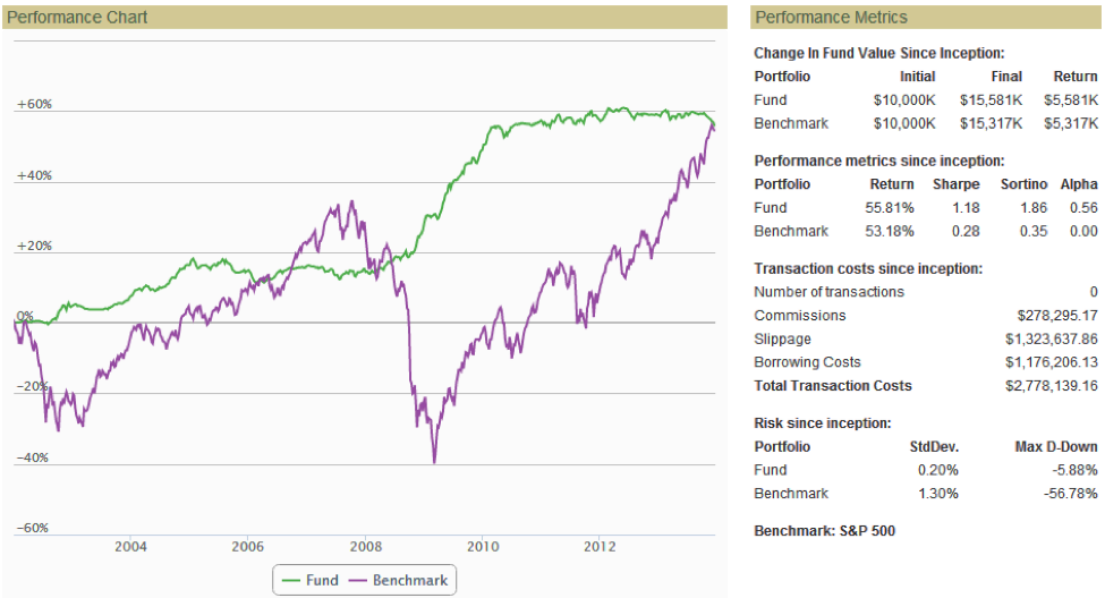

图 B.5　Performance chart for 2002-2013

References

[1] Team Latte. Morgan Stanley and the Birth of Statistical Arbitrage. 2011.

[2] Ganapathy Vidyamurthy. Pairs Trading: Quantitative Methods and Analysi. 2004.

[3] PR Newswire. Knight Capital Group Releases December 2009 Volume Statistics. 2009.

[4] Kerrie McHugh. Bank of America Merril Lynch Unveils New Pairs Algo trading strategy. 2010.

[5] Jesse Eisinger. The hedge fund collapse. 2008.

[6] Hedge Fund Research. Number of Funds. 2009.

综合论文训练记录表

学生姓名	陈炜艺	学号	2010011352	班级	计科 00
论文题目	配对交易——相对价值套利的算法实现与实验分析				
主要内容以及进度安排	主要内容：论文分析一种短期套利策略——配对交易（pairs trading），简言之即寻找历史价格走势相似的股票，当传布（spread）足够宽时卖空走势偏高同时买入走势偏低的股票。该论文目的是引入计算机算法进行实验（程序语言为 Python）构造一个最高历史收益的配对交易策略。同时证明逆势回报有一部分来自人们对公司新闻信息的过激反应，而非价格回归模型的长期相对均衡。 进度安排：前八周进行开发（Developer），后八周进行量化分析（Quant） • 第一周：建立 QSTK 开发环境与熟悉 QSTK • 第二周：投资组合评估与优化的 Python 库建立 • 第三周：事件分析（Event Study）库的建立 • 第四周：市场模拟（Market Simulation）库的建立 • 第五周：时间分析库与市场模拟库的连接 • 第六－八周：财务指标（Financial Indicator）库的建立、连接和测试 • 第九－十周：配对交易数学模型的建立与实现 • 第十一－十三周：利用后台库对配对交易策略进行实验以及分析疑问 • 第十四－十六周：根据实验结果撰写最优策略报告和最终论文 **指导教师签字：** 唐年中 **考核组组长签字：** 李建 2016年 3月 26日				
中期考核意见	论文进行顺利，中期进展顺利 **考核组组长签字：** 李建 2016 年 4月 26日				

指导教师评语	Work mainly done at Geogia Tech.. 指导教师签字： 2014年 6月 4日
评阅教师评语	论文提出了建立了pairing的新的算法. 算法在于传统的方法有较新的突破 评阅教师签字： 2014年 6月5日
答辩小组评语	论文选题研究了好的密码学的一个热点问题. 在利用双线性映射的基础上给出了 一个新的算法.算法设计合理. 答辩小组组长签字： 2014年 6月5日

总成绩：78

教学负责人签字：

2014年 6月16日