FF6120 – Data Science for Decision Making II

Self Introduction

Xie Wenjun （谢文俊）

Email : frankxiewenjun@gmail.com

HP: +65 82280370

Nanyang Technological University

Bachelor in Mathematical Science (Statistics) 2008 – 2012

PhD in Finance 2012 – 2016

Main Research Interest: Actuarial Statistics; Quantitative Finance

Harveston Asset Management

Head, Quantitative Research Department 2016 - now

Course Outline

- 数据科学与金融
  - 数据科学目前的发展与挑战
  - 金融行业的分工和数据科学在其中的应用
  - 量化交易的发展和用到的工具

- Python的基础
  - Python的由来，金融分析中使用Python
  - Python的安装以及常用库的安装
  - IDE种类的介绍以及适合使用的环境
  - 基本数据结构，基础运算以及程序结构

- 金融统计基础
  - Package：NumPy的介绍
  - Package：Matplotlib 数据的可视化
  - Package：Pandas的介绍
  - 数据存储与输出

Course Outline

- 金融统计进阶
  - 概率分布
  - 线性回归
  - 时间序列模型
  - 主成分分析(PCA)
  - 因子分析(Factor Analysis)
  - 风险模型

- 机器学习,分类聚类分析
  - Package：Scikit-learn的介绍
  - Package：Keras的介绍
  - 决策树, KNN, 朴素贝叶斯, K-Means, SVM
  - 人工神经网络在金融中的应用

- 网络金融数据的获取，API以及爬虫
  - Package：BeautifulSoup的介绍
  - Html解析以及网络爬虫
  - API的使用和沟通方法
  - 常用金融数据的收集与获取
  - 基础自然语言处理（NLP）

第一课：数据科学与金融

目录

数据科学的发展与挑战

- 传统数据科学
  - 数据量信息有限
  - 计算机处理模型计算能力有限
- 大数据
  - 互联网的发展
  - 收集数据的手段增多
  - 数据保存的能力提升
- 如何处理大数据
  - 人工智能，机器学习
  - 统计模型

数据科学的发展与挑战

- 市场情绪分析

传统

方式：问卷调查

对象：分析师

处理数据方式：统计百分比，均值，方差等等

大数据

方式：网络数据抓捕，互联网大范围调查

对象：分析师，专业投资者，非专业投资者

处理数据方式：NLP（自然语言处理），情感分析

数据科学的发展与挑战

- 个人信用

传统

方式：特征调查

对象：个人

处理数据方式：根据个人信息，如年龄，职业，收入，婚姻状况等特征来分类。

大数据

方式：互联网数据共享

对象：个人，各类机构

处理数据方式：根据个人过往在各层各面留下的历史数据来进行特征分类。

数据科学的发展与挑战

- 做市交易 （market making）

传统

方式：人工交易

对象：金融产品

处理数据方式：肉眼观察

大数据

方式：量化自动交易

对象：金融产品

处理数据方式：高性能计算机

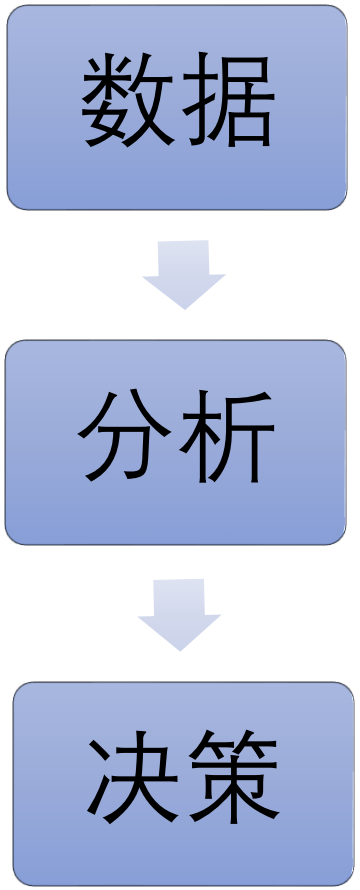
数据科学的发展与挑战

- 面临的挑战

  如何处理数据？（数据质量）
  如何从数据里正确挖掘出我们需要的信息？（数据建模）
  如何确保数据安全？
  数据的架构

金融行业分工及数据科学

- 卖方
  - 投资银行，商业银行，证券公司 ...
- 买方
  - 个人，公司，保险公司，基金，券商资管 ...
- 服务提供商
  - 评级机构，金融研究，数据，系统 ...
- 监管机构，政府机构
  - 央行，银保监会，证监会，政府职能部门



投资银行

- 投资银行业务
  - Primary Market (Pre-IPO financing, IPO, Fund Product, Bond Issuance, M&A …)
- 交易
  - Secondary Market, Structured Products, Proprietary Trading
- 投研
  - Research
- 风险与融资
  - Market Risk, Credit Risk
- 科技
  - System, Trading System, Connectivity

商业银行

- 存贷款业务
  - 居民储蓄，央行借款，贷款 ...
- 担保业务
- 风险

证券公司

- 经济业务
  - 股票交易，债券交易，商品交易 ...
- 承销业务
  - 包销或代销形式帮助发行人发售证券
- 自营业务 （买方）
- 投研
- 风险

基金

- 共同基金 （mutual fund）“公募”
- 对冲基金 （hedge fund）“私募”
- 私募股权基金 （private equity）
- ETF基金 （exchange traded fund）“交易行开放式指数基金”
- 保险基金
- 养老金基金 “主权基金”
- 捐赠 （endowment）

基金

- 基金经理
- 投研
- 交易
- 风险
- 科技与运营

服务提供商

- 评级机构
  - 标普，惠誉，穆迪，Equifax，Transunion
- 投研机构
  - 券商，投行，专业投研机构
- 数据
  - 市场数据 （Market Data）： Bloomberg B-pipe， Thomson Reuters， ACTIV Financial, Interactive Data （ICE）…
  - 基础数据（Fundamental Data）: FactSet, MSCI Barra, IBES (Thomson Reuters), RavenPack，万得，东方财富，朝阳永续 …
  - 另类数据 （Alternative Data）： Quandl …
- 系统
  - 综合系统：Bloomberg Terminal， Thomson Reuters Eikon …
  - 交易系统：Citi Velocity， FlexTrade，通达信，恒生 …
  - 风险管理系统:
  - 服务器：数据中心（Data Center），AWS，阿里云 …

监管机构，政府机构

- 央行 （central bank）
  - 失业率，通胀，汇率，货币基础 …
- 政府监督机构
  - 异常交易，内幕交易，关联账号 …
- 政府职能机构
  - 财政政策，经济指标 …

量化交易

- 主观交易 （discretionary trading）：

- 量化交易（quantitative trading）：

  量化交易是指以数学模型替代人为的主观判断，利用计算机技术从庞大的历史数据中海选能带来超额收益的多种“大概率”事件以制定策略，极大地减少了投资者情绪波动的影响，避免在市场极度狂热或悲观的情况下作出非理性的投资决策。

量化交易

- 量化交易的对象：

股票（Equity）：现货（cash），衍生品（derivatives）

股指（Equity Index）：ETF，衍生品（derivatives）

外汇（FX）：现货（spot），衍生品（derivatives）

债券（bond）：现货（cash），衍生品（derivatives），CDS

商品（commodity）：现货（physical），衍生品（derivatives）

利率（interest rate）：衍生品（derivatives）

量化交易

- 量化交易策略种类:

股票（Equity）：

-价值投资
-多因子选股（multi-factor model）
-机器学习

CTA （Commodity Trading Advisor）：

-趋势策略（trending strategy）
-回归策略（mean reversion strategy）
-统计套利 （statistical arbitrage，pairs trading）
-波动性策略 (volatility strategy)

高频 （High Frequency Trading）

-做市（ market making ）
-套利（arbitrage）
-短期区趋势（ directional trading ）

量化交易

- 多因子选股模型（multi-factor model）：

  多因子模型是量化选股中最重要的一类模型，其基本思想就是找到某些和收益率最相关的指标。并根据该指标，构建一个股票组合，期望该组合在未来的一段时间跑赢指数。

  因子种类

  市场因子：市场风险，波动性，流动性，换手率…

  风格因子：

  - 价值因子：PE，PB …
  - 成长因子：EPS增长率，营业利润增涨率…
  - 规模因子：总市值，流通市值…

  财务因子：EPS，每股净资产，流动比率，ROE …

  技术因子：动量指标，MACD，RSI …

量化交易

- 多因子选股模型（multi-factor model）：

  多因子选股流程

  数据预处理：基础数据采集，离群值处理，数据标准化

  单因子检验：特征分析，中性化处理，回归法分析，分层回测

  大类因子合成：因子相关性分析，细分因子合成

  模型构造：确定因子权重个股打分/回归分析

  组合优化：添加约束条件，二次规划

  常见问题

  过度拟合

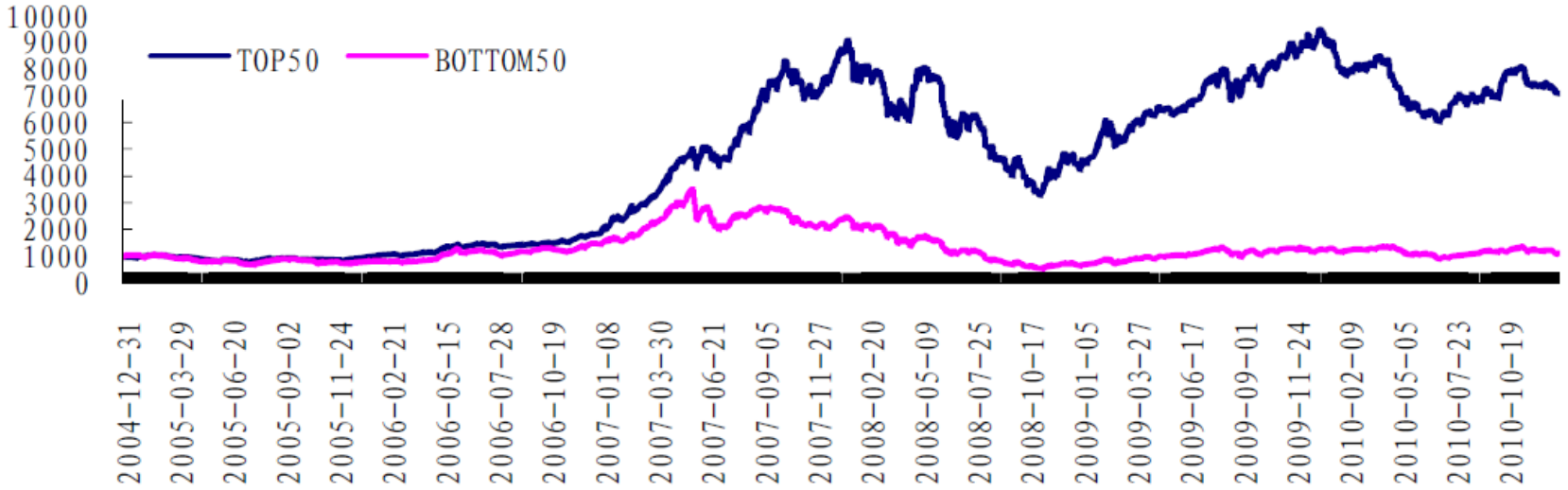
  因子相关性过高，多重共线性

量化交易

- 多因子选股模型（multi-factor model）：

来源：安信证券研究中心

多因子选股打分法：

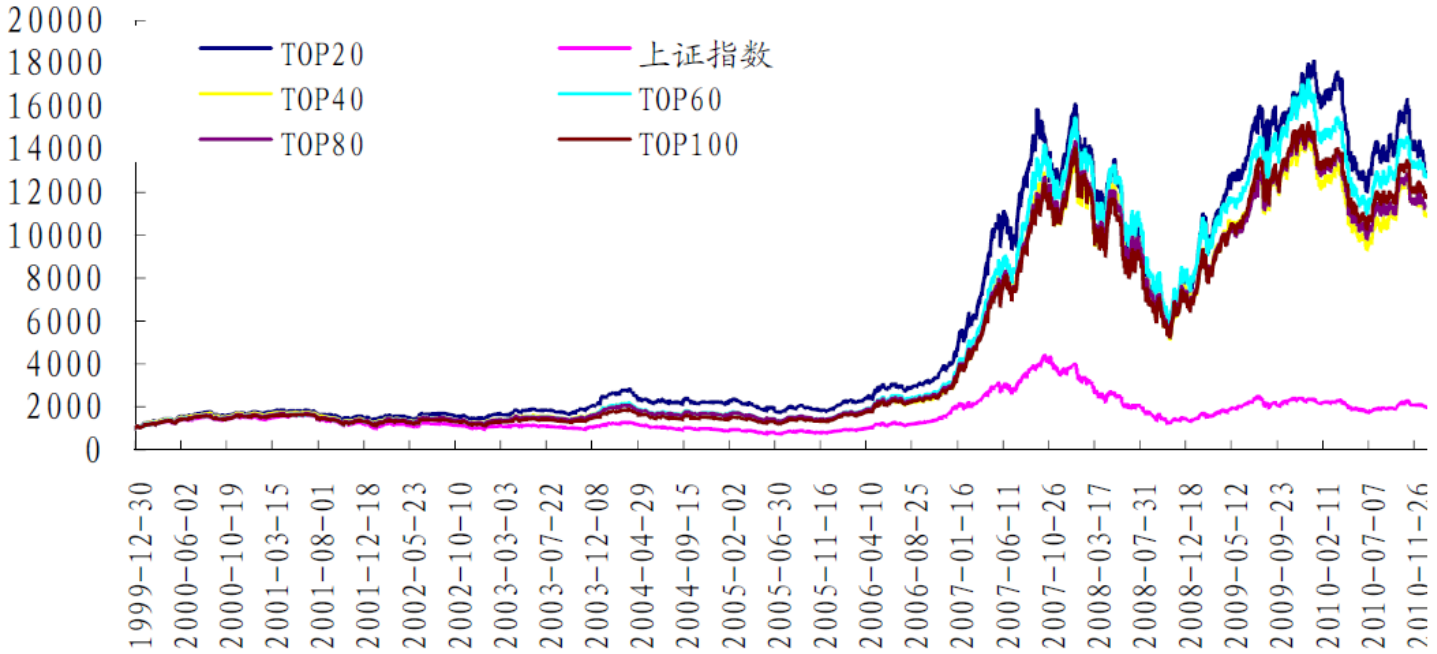


量化交易

- 多因子选股模型（multi-factor model）：

来源：安信证券研究中心

多因子选股打分法：



量化交易

- 多因子选股模型（multi-factor model）：

  多因子选股发展与创新

  新因子的发现：大数据下的情绪因子

  方法的创新：动态调整，机器学习

量化交易

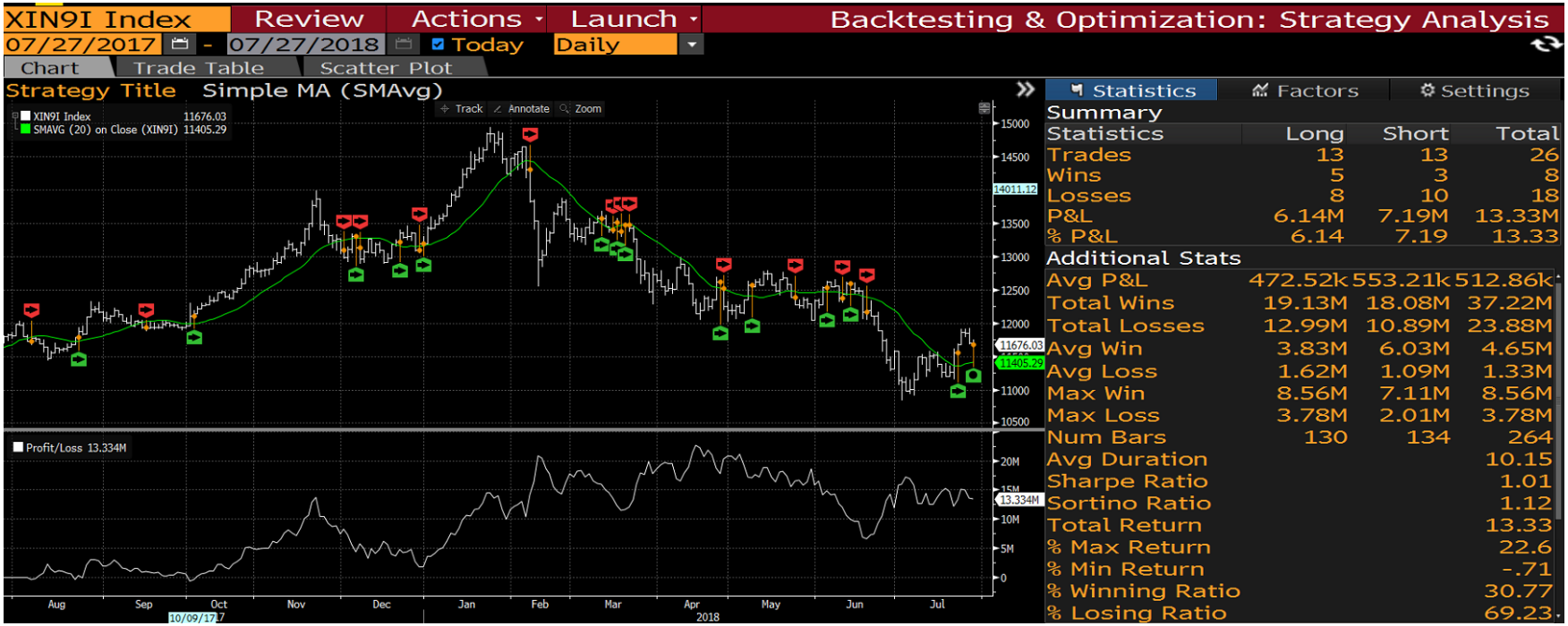
- CTA趋势策略:

Simple MA策略

量化交易

- CTA趋势策略:

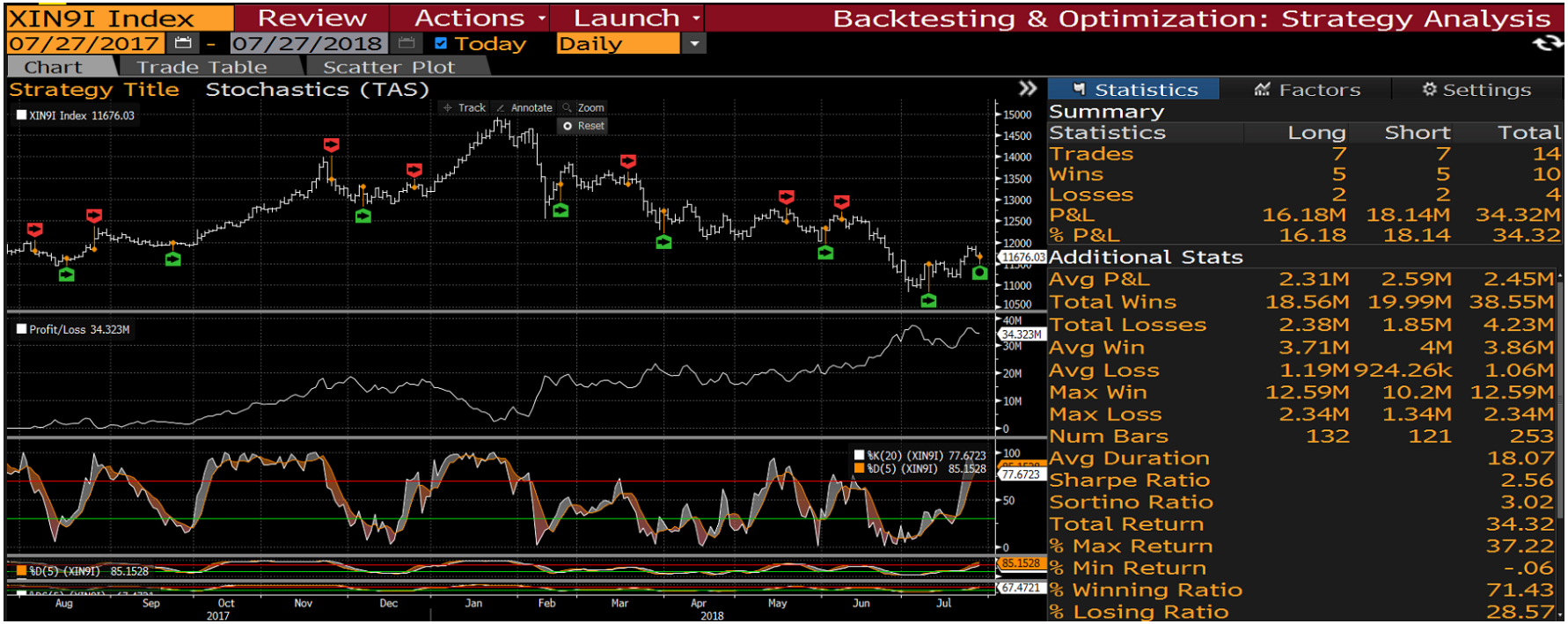
Simple MA策略



量化交易

- CTA回归策略:

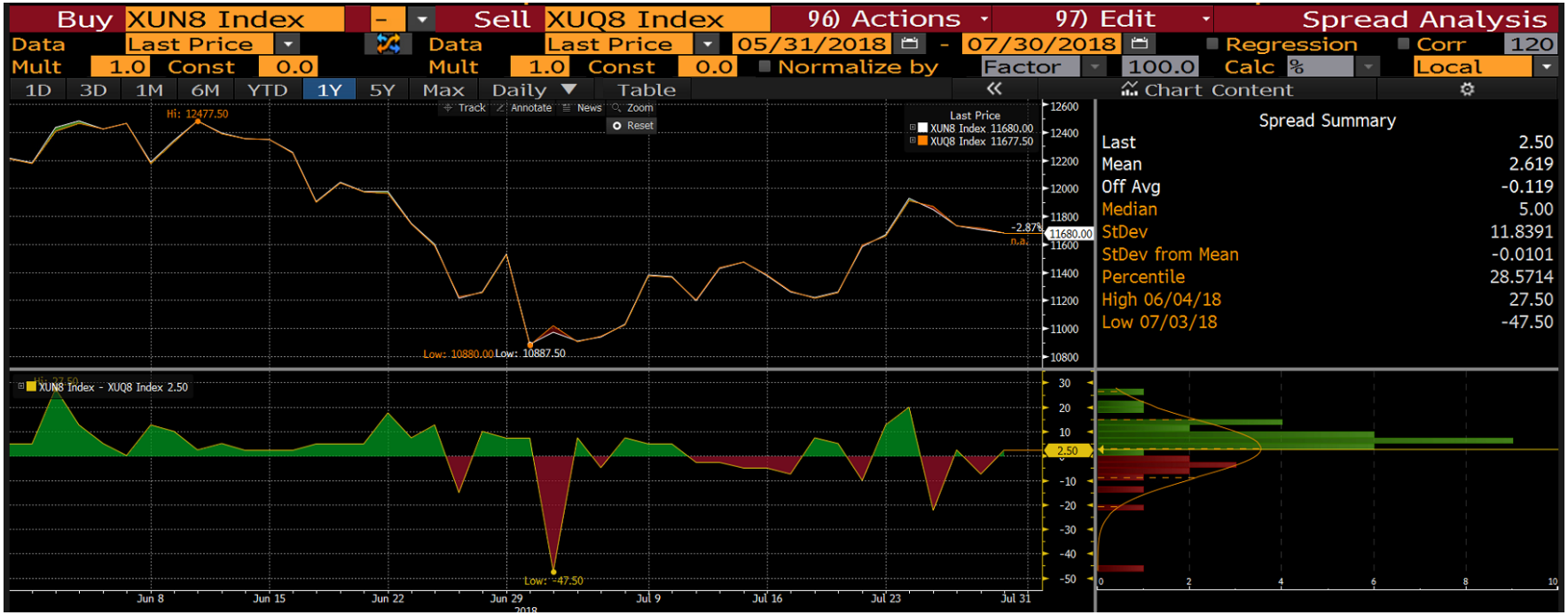
Stochastic策略



量化交易

- CTA统计套利：

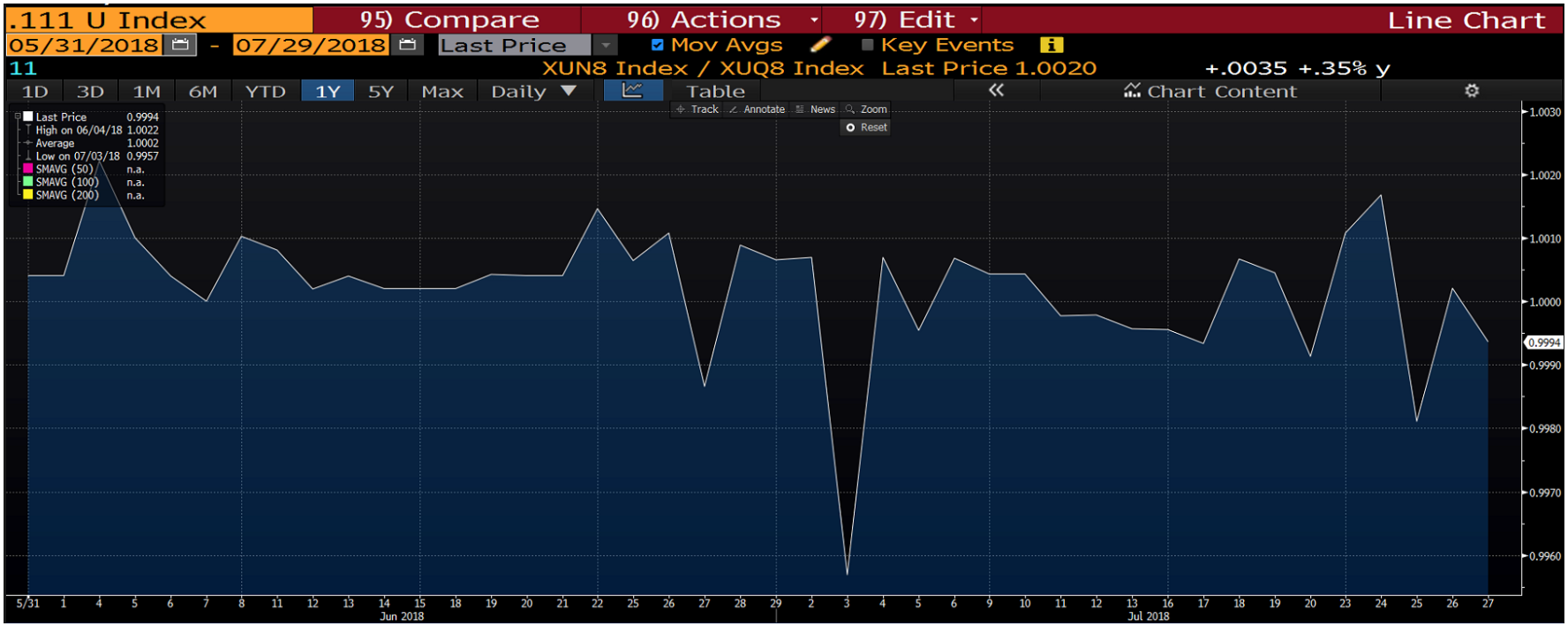
Pairs Trading Distance Method （A - B）



量化交易

- CTA统计套利：

Pairs Trading Cointegration Method （A/B）



量化交易

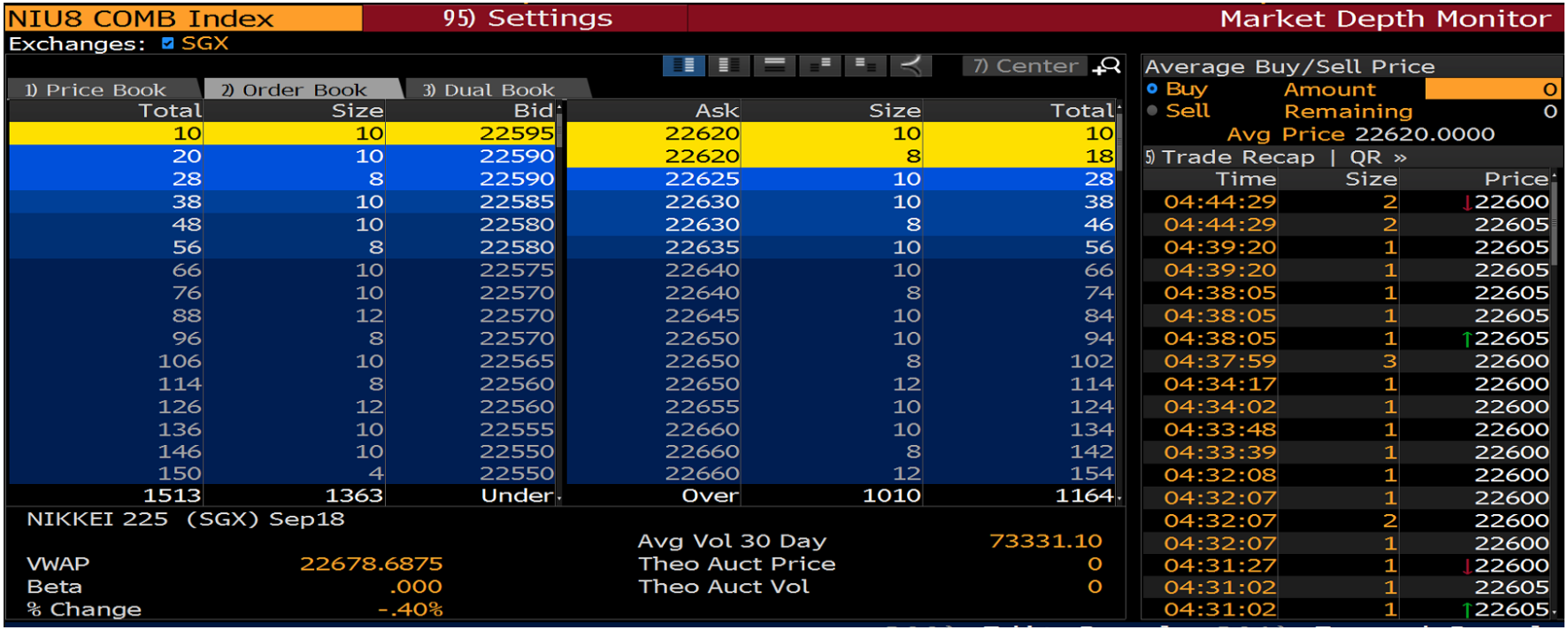
- HFT套利：

同一产品在不同交易所之前的套利

量化交易

- HFT做市：

Liquidity Maker vs. Taker

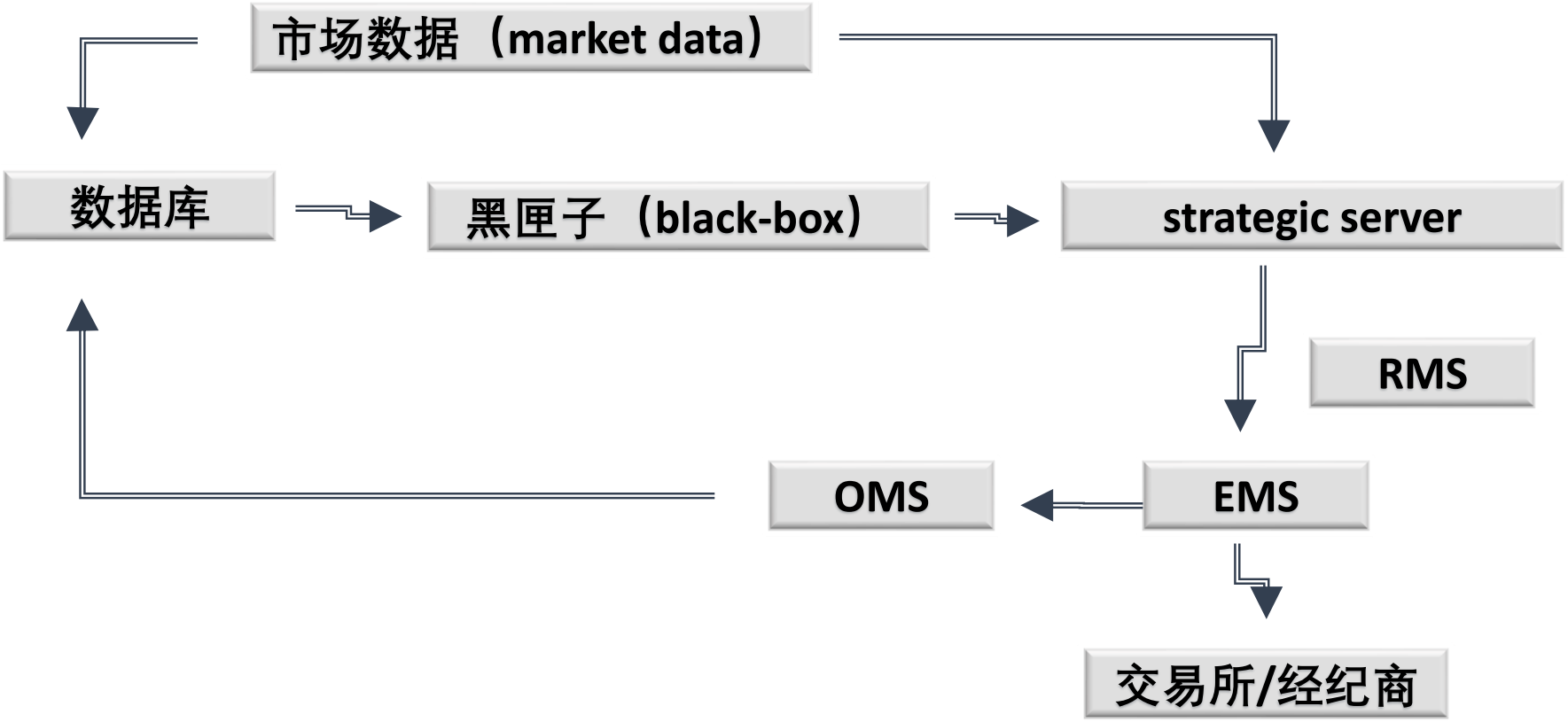


量化交易

- 量化交易的组成部分：
  - 数据库 （historical database）
  - 市场数据 （market data, tick data）
  - 策略 (黑匣子，black-box)
  - Strategic Server
  - EMS (execution management system)
  - OMS (order management system)
  - RMS (risk management system)
  - connectivity

量化交易

- 量化交易的Architect：



编程语言（programming）

- 主流科学编程软件：

MatLab； R； Python； Octave； Eviews； SAS； STATA； SPSS

- 数据库

File-Based； SQL； DB2； Oracle； Cassandra； Kx System（financial time series database， Q语言）

BI工具：Tableau、Qlikview、FineBI

- 底层：

C/C++;

编程（programming）

- 编程逻辑能力
- 解决问题的能力
- 适应新技术的能力
- 实践

Group Project

- Description

  This project requires a replication of an existing study on financial data analysis using Python. Students can either choose the papers provided or any other academic study of their own interest related to financial data analysis.

- Team Formation

  3 ~ 4 students per group

Group Project

- Requirement
  - The main methodology part of the study is expected to be replicated, though some simplifications can be made.
  - You can choose any instrument or market that is of your interest to conduct the study.
  - Potential data source can be Bloomberg, Yahoo Finance, Choice or any other reliable data source you can find.
  - Additional analysis beyond the paper is welcomed.

Group Project

- Expected Deliverables
  - A report including the following sections: introduction, methodology (simplified), data description, analysis of the result, conclusions.
  - A .py file containing the codes that you conduct the analysis.

- Assessment
  - Correct understanding of the methodology of the chosen paper
  - Application of the appropriate Python packages to conduct relevant studies
  - Data collection and cleaning
  - Analysis of the results
  - Writing of the report

Group Project

- Deadline:

Online submission: 11:59 pm 9th Sep

Hard-copy submission: 6:00 pm 10th Sep

数据收集

- Bloomberg Data Downloading Through Excel:

https://www.youtube.com/watch?v=XZGs4AqdxeE&t=20s

- Bloomberg Terminal:

https://www.youtube.com/watch?v=LE8HiHZcgEE