上证 50 股票聚类实证分析

这一部分主要通过 OLB 算法对上证 50 中的 50 支股票进行聚类分析。其中我们提出了新型的 OLB 方法对股票进行聚类。该部分主要包括如下几部分：

- 数据的导入与预处理
- OLB 算法基本原理
- 实证结果

一、导入数据与预处理

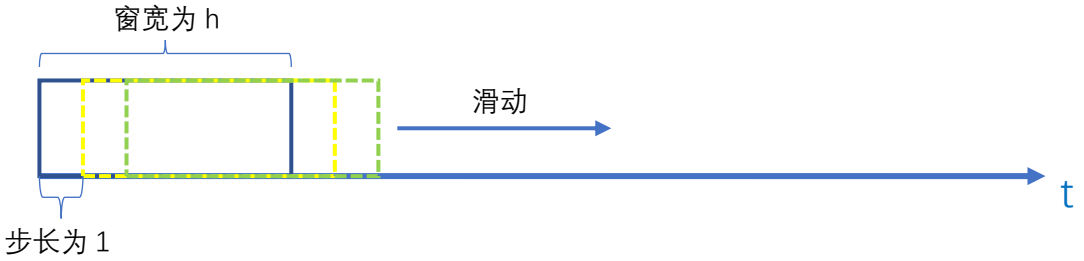
1.将上证 50 的 50 只股票导入，剔除缺失值较为严重的三支：**闻泰科技，中泰证券，中金公司**，共选择有 7 项指标：**收盘价、成交量、换手率、市盈率、市净率、市销率、市现率**。即原始数据为 47 个七维时间序列，序列长度为 728。

2.对原始数据进行 **Z-标准化**，并对个别缺失值采用**线性插值法**进行填充。

3.选取窗宽 h 为 20，对 47*7=329 条一维序列分别进行断点查找，得到断点所构成的七维时间序列，以进行后续聚类操作。

二、OLB 聚类

多维时间序列的聚类方法本文选用 **OLB 聚类算法**。在给定基准序列X_t与待比较序列$\{Y_t^1, Y_t^2, \cdots, Y_t^n\}$的条件下，通过 OLB 算法查找距离基准序列$X_t$最近的序列$Y_t^*$。该算法的原理为选定一个给定窗宽为 h 的滑窗，在时序时间轴上以步长为 1 进行滑动，使之遍历各个时间序列。



在每个滑窗内，分别计算待比较序列Y_t^i与基准序列X_t之间的距离，并查找出距离最近的序列Y_t^*。在滑窗遍历整个时间轴后，统计各个待比较序列$Y_t^i = Y_t^*$的频数，作为待比较序列与基准序列间的距离刻画。

该算法的具体内容如下所示：

基于滑动窗口的 ATS 算法
输入：基准多维序列X_t，待比较多维序列$\{Y_t^1, Y_t^2, \cdots, Y_t^n\}$，窗宽为 h 输出：距离$X_t$最近的序列$Y_t^*$ 1.取窗宽下界起始点 start=0，计算序列长度 length；

2.将各序列进行标准化，并通过基于滑动窗口的 ATS 算法进行变点查找；
3.将得到变点序列以更新原始序列X_t，$\{Y_t^1, Y_t^2, \cdots, Y_t^n\}$；
4.定义待比较序列的键值对 dict=$\{Y_t^i: 0,\ i = 1,2,\cdots,n\}$，键为待比较序列的名称$Y_t^i$，所对应值 value 初始化为 0；
5.当 start+h<=length 时，执行
6. 定义相关性距离向量 P_lst=[ ];
7. 遍历 待比较序列$\{Y_t^1, Y_t^2, \cdots, Y_t^n\}$中的$Y_t^i$:
8. 计算X_t与Y_t^i的相关性距离P_{XY_i};
9. 将P_{XY_i}存入 P_lst
10. 计算出 P_lst 中的最小值，并查找其所对应的Y_t^i，赋值 dict 中的键对应值 value = value+1;
11. start = start+1;
12. 结束
13.计算 dict 中 value 最大值所对的键名，即距离X_t最近的序列Y_t^*，输出Y_t^*;
14.结束

聚类分为两次：第一次聚类投入全部 7 项指标，即对七维时间序列进行聚类；第二次聚类仅投入收盘价，即对一维时间序列进行聚类。OLB 聚类中所选窗宽 h 为 20。

聚类的具体结果见 word 文档【上证 50 聚类结果（全部指标）】与【上证 50 聚类结果（仅收盘价）】

现仅讨论中国石化，上海机场，浦发银行三只股票的聚类结果。

第一次聚类中，三只股票的结果如下所示：

基准序列	中国石化	上海机场	浦发银行
距离其最近的序列	中国石油	华泰证券	光大银行
距离其最远的序列	海天味业	工商银行	海螺水泥

其时序图如下图所示（仅展示收盘价），其中第一行为基准序列，第二行与第三行分别为距离其最近和最远的序列。

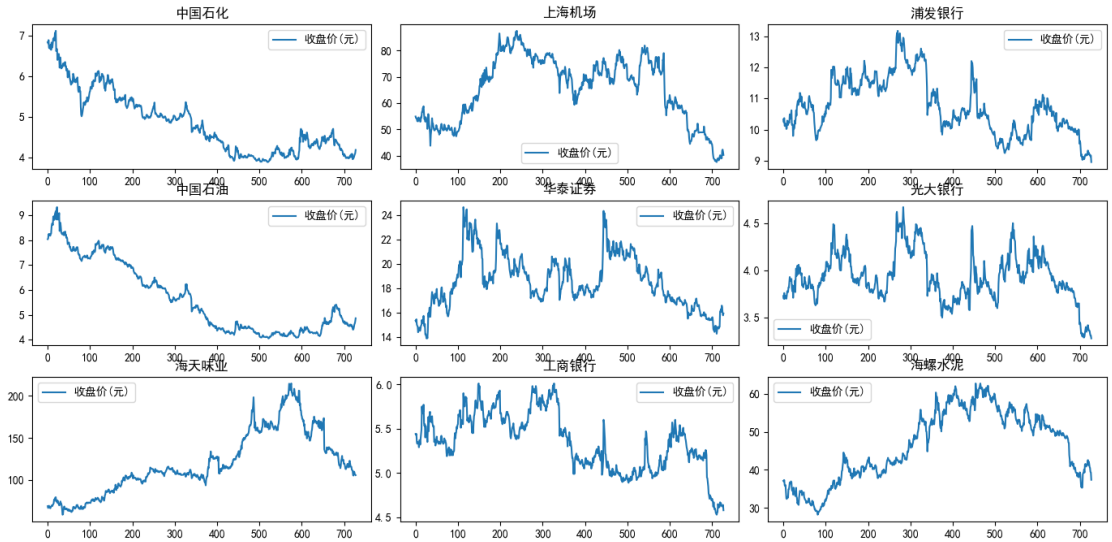


Figure 10

第二次聚类中，三只股票的结果如下所示：

基准序列	中国石化	上海机场	浦发银行

距离其最近的序列	中国石油	中国平安	农业银行
距离其最远的序列	万华化学	恒生电子	华泰证券

其时序图如下图所示（仅展示收盘价），其中第一行为基准序列，第二行与第三行分别为距离其最近和最远的序列。

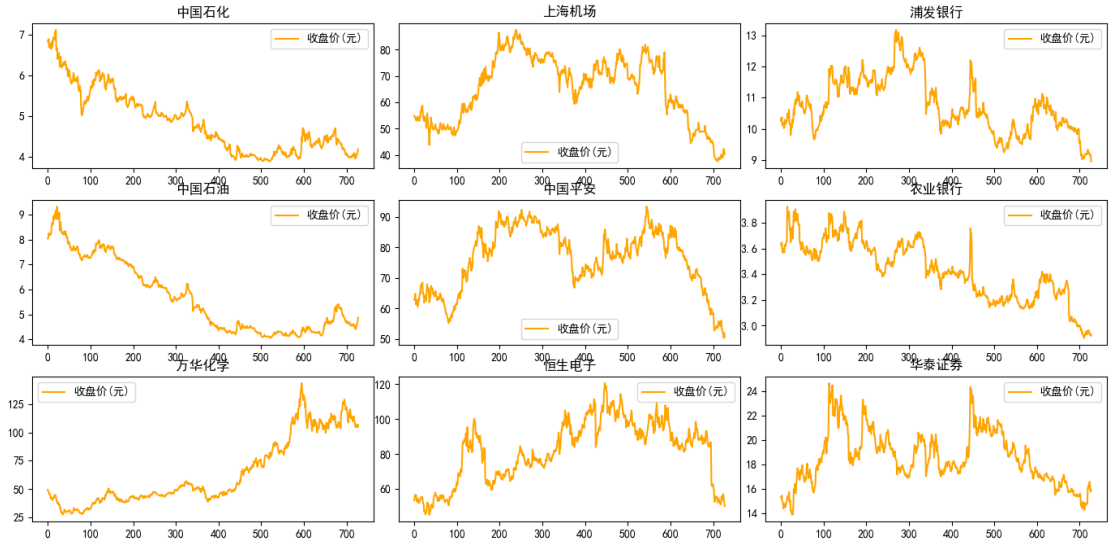


Figure 9