JM因子研究报告 童屹邦

提交的因子如下，包含两个范例与其他4个因子：

1. midpiceRtn，示例因子A
2. orderImbalance，示例因子B
3. pricejump，因子1
4. max\_portion\_orthogonal，因子2
5. voi，因子3
6. z，因子4

（若需要对样本外数据进行测试，请将level1数据和最优十笔委托的数据分别放在两个文件夹中，并在main\_demo的第12、13行的read\_data中的文件夹改为要测试数据的文件夹即可。（数据文件夹和main\_demo在同一路径下））

**代码框架：**

Main\_demo.py：主运行函数，会调用以下三个文件

Data\_Process.py：一些数据处理函数

factors.py：因子

performance\_analysis：包含了一些因子分析需要的函数

对于给定的四个月的数据，全部因子全部运行一遍的总运行时间大约为3分钟。

**数据处理：**

在计算因子时，剔除了每次开盘前15秒、收盘最后15秒的数据。主要是考虑到开盘最开始的几个tick有缺失值，以及一些因子需要构造滑动平均，收盘同理。以夜盘为例，因子值生成的时间区间为21:00:15到22:59:45。另外几次开盘的时间同理。

**因子结果：**（所有结果都在**result文件夹**中）

**IC值**：

midpriceRtn -0.062134

orderImbalance 0.079317

pricejump 0.042017

max\_portion\_orthogonal -0.043135

voi 0.076434

z -0.091556

**因子间相关性**：详见”Correlation\_matrix.txt”，最高的一组相关性为0.25

**头尾quantile**：详见”quantile均值.txt”

ACF、因子统计分布均保存在result文件夹下面的每个因子的文件夹中

最后五天的因子值矩阵保存在csv中

**因子1：pricejump**

计算方法：

**Step1.** 计算当前midprice与max(过去20tick的midprice)的差值，得到序列x

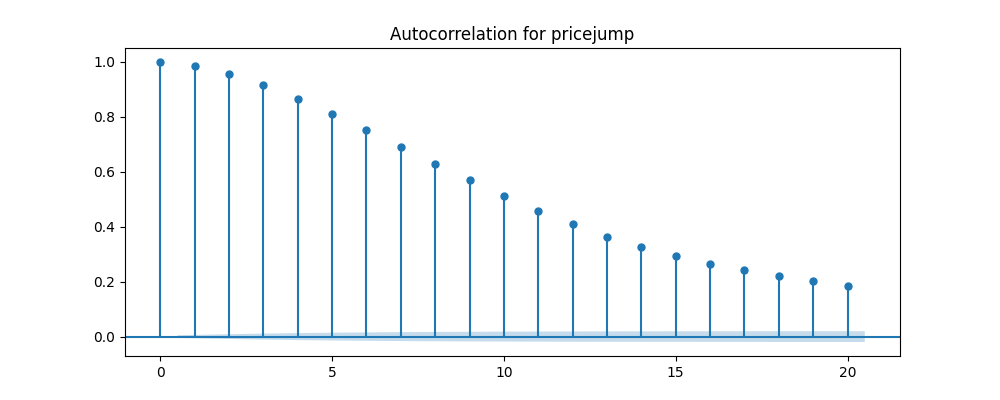
**Step2**. 计算序列x的EMA(span=15)，得到序列y

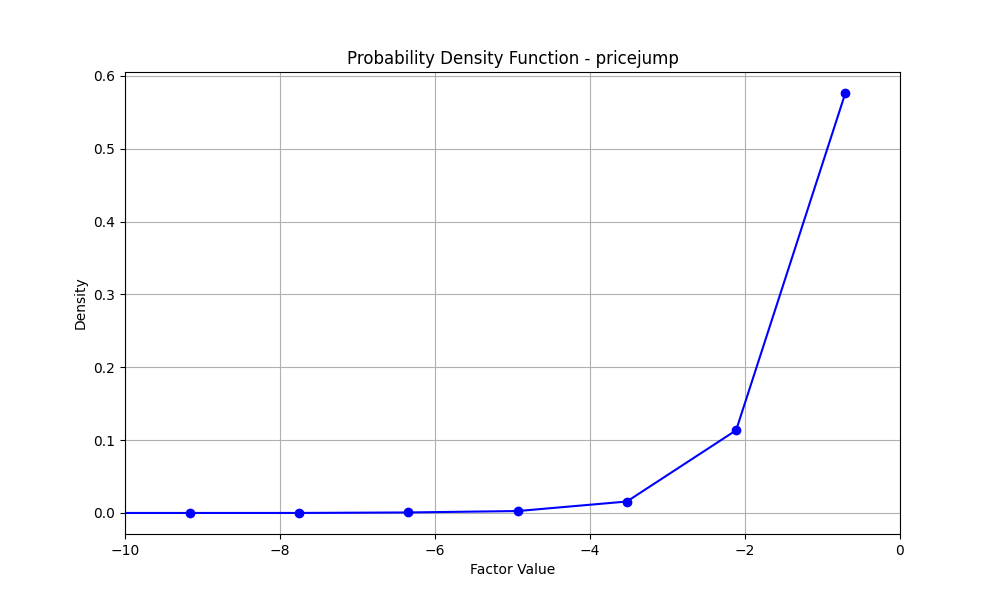
**Step3.** 对序列y进行shift(1)，即当前tick用的是上一个tick的因子值，得到最终因子值。

**解释**：最开始只进行了step1，逻辑是检测价格在短时间内的显著变动。该因子x在单日的表现已经足够优秀，有些日期能达到0.09，试了几个日期都能在0.04以上；但是回测所有时间段的话IC就降到了0.01。所以或许该因子对于不同市场情况的鲁棒性不够强，因此考虑对其进行滑动平均降噪。

一开始尝试了5个tick的MA，IC能到0.03以上，corr也很低，但quantile的头尾差异不明显；

然后尝试了span=10的EMA，IC也能到0.03以上，quantile的差异也比较显著，不过corr稍微有点大，和midpiceRtn有0.58。应该是EMA赋给当前与上一个tick的权重较大造成的，这和midpiceRtn使用数据太相似。 （或许可以正交化试试？）

 最后尝试对上面的EMA进行shift(1)，即当前tick的EMA值使用上一个EMA的值。这样可以避免使用当前tick的midprice数据，理论上可以降低其与midpiceRtn因子的corr。事实确实如此，IC上升到了0.42，和其余因子之间的corr都很低，quantile表现尚可。因此就使用这样的方法来构造因子。



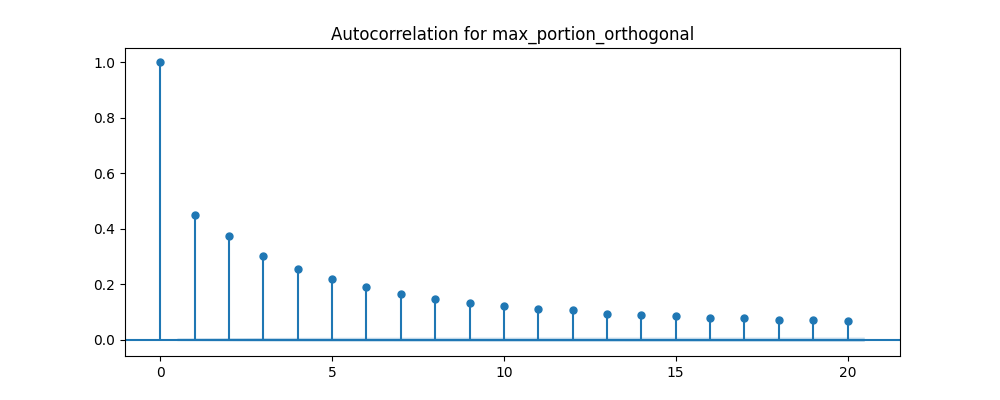
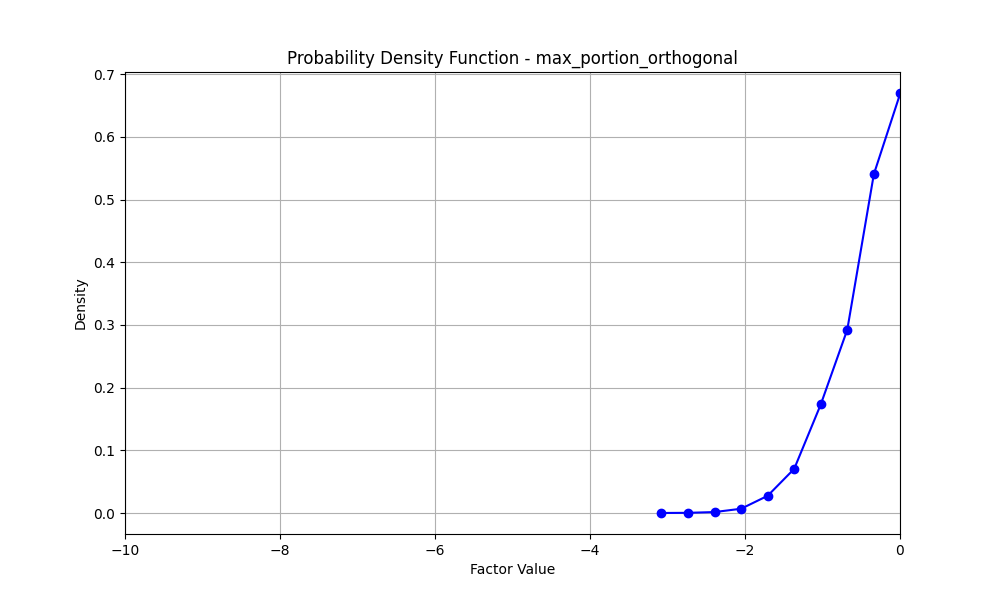
**因子2：max\_portion\_orthogomal**

计算方法：

**Step1**. 计算最优价前十笔委托中最大的一个quantity占总quantity的比重。对Bid和Ask分别计算，得到Bid的比重和Ask的比重

**Step2**. Log(Bid的比重/Ask的比重)

该因子IC有-0.078，quantile的头尾区分也很显著。然而其和orderImbalance的corr有0.52。这确实是很明显的，因为其构造和orderImbalance非常相似。因此我考虑将该因子对orderImbalance正交化，得到max\_portion\_orthogonal。

正交化后的因子IC有-0.043，表明该因子确实有其他独特的、有效的信息。不过quantile头尾部的均值倒是区分不太明显。因此这里正交有利有弊，确实降低了corr，但也降低了quantile的头尾显著性。

（注：我对最优价前十笔委托的数据尝试了很多种构造方法，构造出IC不错的因子基本都和orderImbalance以及max\_portion\_orthogomal这两个因子相关性很高。 可能这种纯用Volume/Quantity数据构造的因子的信息都是类似的？）

**因子3：订单失衡voi因子**

来源：20200709-中信建投-高频量价选股因子初探

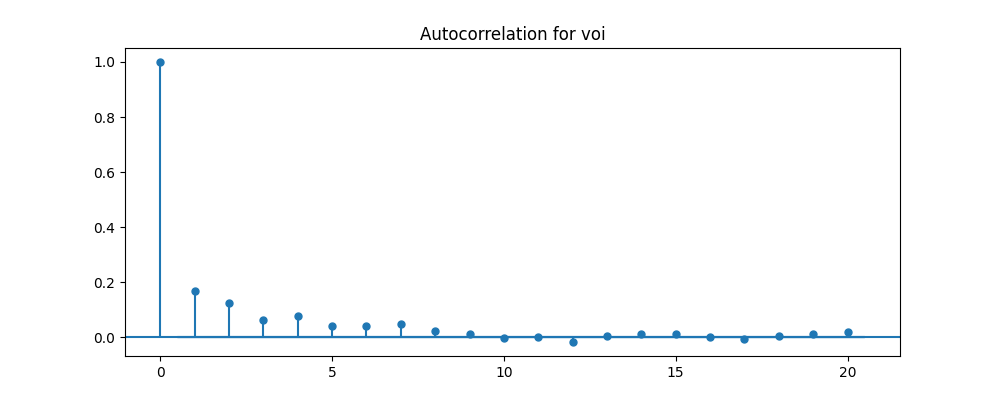
计算方法：

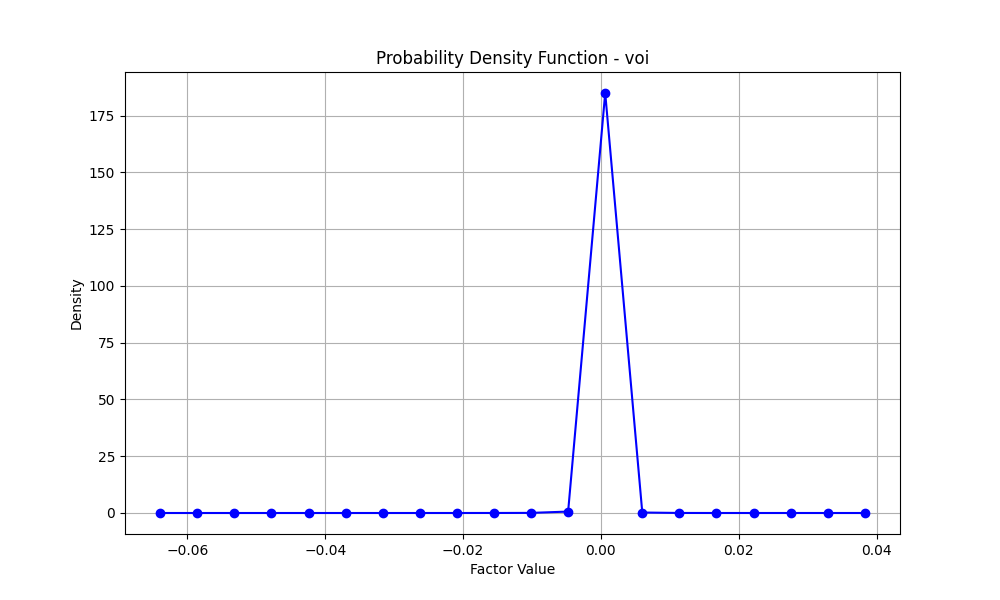
**Step1**. 计算BidPrice1、BidVolume1、AskPirce1、AskVolume1与上一时刻的变化量

**Step2**. 对于BidVolume1的变化量，如果BidPrice1上升，则不变；如果BidPrice1下降或保持不变，则将其设为 0。对于AkVolume1的操作同理

**Step3**. (BidVolume1变化量-AskVolume1变化量)/data['Volume']，即为因子值

该因子IC值为0.076，与其他因子的相关性都较低（皆低于0.2）。quantile的头尾分布也较为优秀。



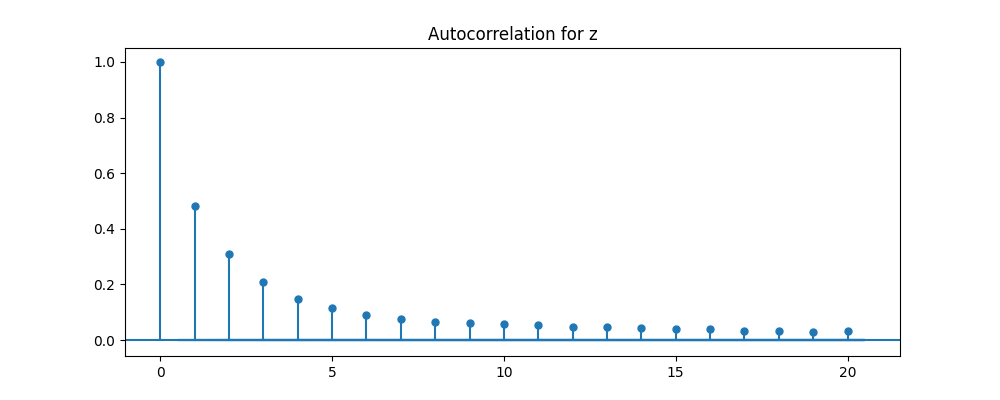
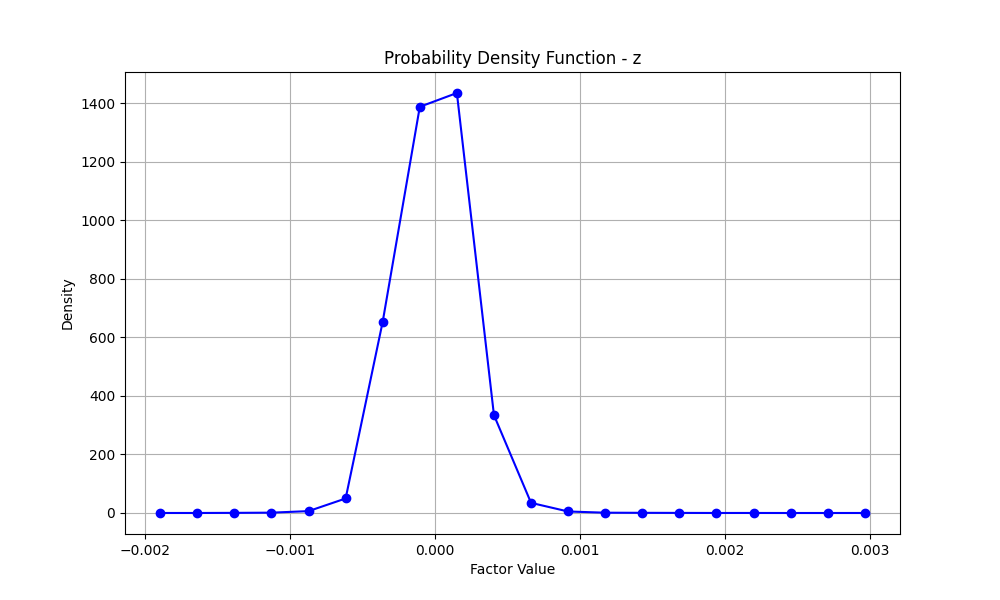


**因子4：z因子**

来源：20130904-民生证券-指令单薄与指令单流——资金交易策略之四

**计算方法:log（Midprice - LastPrice）**

该因子IC值为-0.09，与其他因子的相关性都较低（最高0.25）。quantile的头尾分布也非常优秀。

****

其余尝试过的因子还包括：高频偏度与峰度、量价相关性（海通证券），效果一般