北京大学
PEKING UNIVERSITY

高频交易价格研究

北京大学数学科学学院

王瑞辰 郝思越 陈奕涵 王彦喆

2022年12月5日

1 理论背景

2 价格估计量是否随机

3 价格估计量能否预测交易价格

4 总结与展望

根据高频数据预测未来价格

- 订单簿：包含这一时刻未完成的订单中的所有信息。
  - Bid-price:买方报价，其中最高的报价为P^b
  - Bid-size:买方报价的数量，报价P^b的数量为Q^b
  - Ask-price:卖方报价，其中最低的报价为P^a
  - Ask-size:卖方报价的数量，报价P^a的数量为Q^a
  - Spread:价差，P^a-P^b
  - Imbalance:$\frac{Q^b}{Q^a+Q^b}$，衡量最优买卖报价的数量比重，简记为I

- 两种对未来价格的预测：
  - Mid-price:用最优买卖报价的均值估计未来价格，即$\frac{P^a+P^b}{2}$
  - Weighted mid-price:使用Imbalance加权估计未来价格，即$IP^a + (1-I)P^b$

一种新的未来价格预测

- Micro-price:高频数据条件下使用历史数据估计未来价格
  - 是一列价格的极限值（如果存在）
  - 之后第i次Mid-price改变时刻的Mid-price对于当下时刻的订单簿信息的条件期望作为第i个价格数据

$$P_t^i = \mathbb{E}[M_{\tau_i} \mid \mathcal{F}_t].$$

- 其中的M是Mid-price，下角标指的是Mid-price的第i次改变时刻，即

$$\tau_1 = \inf\{u > t \mid M_u - M_{u-} \neq 0\}$$

$$\tau_{i+1} = \inf\{u > \tau_i \mid M_u - M_{u-} \neq 0\}$$

前提假设与计算方式

- 假设1：$\mathcal{F}_t = \sigma(M_t, I_t, S_t)$代表着订单簿上信息总和，其转移概率满足马氏链的条件

$$\mathcal{F}_t = \sigma(M_t, I_t, S_t)$$

- 假设2：Mid-price的变动值与改变前mid-price的值独立

$$\begin{aligned}&\mathbb{E}[M_{\tau_i} - M_{\tau_{i-1}} \mid M_t = M, I_t = I, S_t = S] \\ &\quad = \mathbb{E}[M_{\tau_i} - M_{\tau_{i-1}} \mid I_t = I, S_t = S], \quad t \le \tau_{i-1}.\end{aligned}$$

- 我们得到第i个价格序列的计算方式：

$$P_t^i = M_t + \sum_{k=1}^{i} g^k(I_t, S_t)$$

- 其中

$$g^1(I, S) = \mathbb{E}[M_{\tau_1} - M_t \mid I_t = I, S_t = S]$$

$$g^{i+1}(I, S) = \mathbb{E}[g^i(I_{\tau_1}, S_{\tau_1}) \mid I_t = I, S_t = S], \quad \forall i \ge 0$$

理论计算方式

- 有关计算的证明如下：

$$
\begin{aligned}
P_t^1 &:= \mathbb{E}[M_{\tau_1} \mid \mathcal{F}_t] \\
&= \mathbb{E}[M_t \mid M_t, I_t, S_t] \\
&\quad + \mathbb{E}[M_{\tau_1} - M_t \mid M_t, I_t, S_t] \quad \text{Assumption 1} \\
&= M_t + \mathbb{E}[M_{\tau_1} - M_t \mid I_t, S_t] \quad \text{Assumption 2} \\
&= M_t + g^1(I_t, S_t) \quad \text{Definition}
\end{aligned}
$$

$$
\begin{aligned}
P_t^i &:= \mathbb{E}[M_{\tau_i} \mid \mathcal{F}_t] \\
&= \mathbb{E}[M_t \mid M_t, I_t, S_t] \\
&\quad + \sum_{k=1}^{i} \mathbb{E}[M_{\tau_k} - M_{\tau_{k-1}} \mid M_t, I_t, S_t] \quad \text{Assumption 1} \\
&= \mathbb{E}[M_t \mid M_t, I_t, S_t] \\
&\quad + \sum_{k=1}^{i} \mathbb{E}[M_{\tau_k} - M_{\tau_{k-1}} \mid I_t, S_t] \quad \text{Assumption 2} \\
&= M_t + \sum_{k=1}^{i} g^k(I_t, S_t) \quad \text{Definition}
\end{aligned}
$$

$$
\begin{aligned}
g^{i+1}(I, S) &= \mathbb{E}\left[M_{\tau_{i+1}} - M_{\tau_i} \mid I_t = I, S_t = S\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[M_{\tau_{i+1}} - M_{\tau_i} \mid S_{\tau_1}, I_{\tau_1}\right] \mid I_t = I, S_t = S\right] \\
&= \mathbb{E}\left[g^i(I_{\tau_1}, S_{\tau_1}) \mid I_t = I, S_t = S\right] \text{ for } i \geq 1
\end{aligned}
$$

计算Micro-price

- 统计在每组状态下（一个状态指的是Imbalance和Spread的一种二元组合）Mid-price change的转移概率
- Imbalance的个数取为n=5种：0，0.25，0.5，0.75，1
- 我们将（0.74，1]视为1，将（0.59，0.74]视为0.75，将（0.41，0.59]视为0.5，将（0.26，0.41]视为0.25，将[0,0.26]视为0
- 需要使用当日的数据估计在所有的状态下的Mid-price change的概率，所以我们需要每一种状态均出现过且数量不能过于小，否则会出现分母为0的情况或者由于数据太少导致估计出的转移概率并不准确
- 我们将调整Spread的取值方式使得估计出的转移概率尽可能准确

计算Micro-price

- 如果spread超过0.01的数量不大于全部数据量的1/10，那么我们认为spread几乎没有影响，spread仅分为一组，也就是总计1*5=5种状态
- 若否，如果spread超过0.02的数目不大于全部数据量的1/10，那么我们将spread分为2组，等于0.01的为一组，超过0.01的为一组，此时有10种状态
- 若仍否，我们按照3分位点将spread分为三组，此时有15种状态
- 对于后两种情况，如果仍然出现了某种状态的个数为0，那么我们就将spread的个数减少1，直至只剩下一种Spread，如果此时还有某种状态的个数为0，那么说明imbalance出现不够5种的情况，属于极少的特殊情况，我们选择删去此股票这一天的数据。
- 由此Spread的取值不是固定的，m=1或2或3

计算Micro-price

- 而对于Mid-price change的取值，对于一些价格较高的股票来说，选择$\{-0.01，-0.005，0.005，0.01\}$作为中间价格变动值并不合适
- 取本日内所有的非零的Mid-price change的绝对值数据的20%分位数，50%分位数和80%分位数，如果互不相同，我们选择0.2分位数和0.8分位数以及它们的相反数作为Mid-price change的四个值
- 否则使用$\{-0.01，-0.005，0.005，0.01\}$作为中间价格变动值

计算Micro-price

- R（n*m,4）：r_{ij}是状态为i，Mid-price change为j的数据占状态为i的数据的比例
- Q（n*m,n*m）：q_{ij}的数值是Mid-price change为0且状态从i转移至j的数据占状态为i的数据的比例
- T（n*m,n*m）：t_{ij}的数值是Mid-price change不为0且状态从i转移至j的数据占状态为i的数据的比例

$$R_{xk} := \mathbb{P}(M_{t+1} - M_t = k \mid X_t = x)$$

$$Q_{xy} := \mathbb{P}(M_{t+1} - M_t = 0 \wedge X_{t+1} = y \mid X_t = x)$$

$$T_{xy} := \mathbb{P}(M_{t+1} - M_t \neq 0 \wedge X_{t+1} = y \mid X_t = x)$$

计算Micro-price

- 估计出三个转移概率矩阵可以用来计算Micro-price

$$
\begin{aligned}
G^1(x) &= \mathbb{E}[M_{\tau_1} - M_t \mid X_t = x] \\
&= \sum_{k \in K} k \cdot \mathbb{P}(M_{\tau_1} - M_t = k \mid X_t = x) \\
&= \sum_{k \in K} \sum_{u} k \cdot \mathbb{P}(M_{\tau_1} - M_t = k \wedge \tau_1 - t = u \mid X_t = x)
\end{aligned}
$$

$$
G^1(x) = \left(\sum_s Q^{s-1} R \right) K = (1 - Q)^{-1} RK
$$

$$
G^{i+1}(x) = \left(\sum_s Q^{s-1} T \right) G^i(x) = (1 - Q)^{-1} T G^i(x)
$$

$$
B := (1 - Q)^{-1} T
$$

$$
P_t^i = M_t + \sum_{k=0}^{i} B^k G^1
$$

计算Micro-price

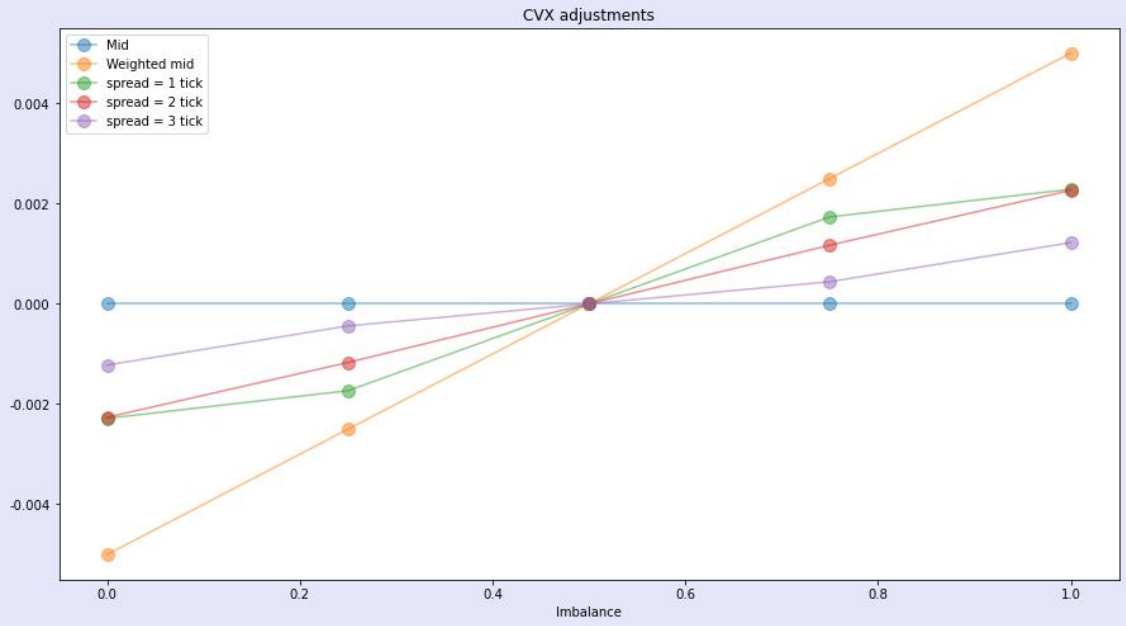
对数据进行复制，已有的一组$(I_t, S_t, I_{t+1}, S_{t+1}, dM)$，再得到一组$(1 - I_t, S_t, 1 - I_{t+1}, S_{t+1}, -dM)$，这样的复制操作加上Mid-price change的对称式选择方法可以保证得到的价格序列是收敛的，即：

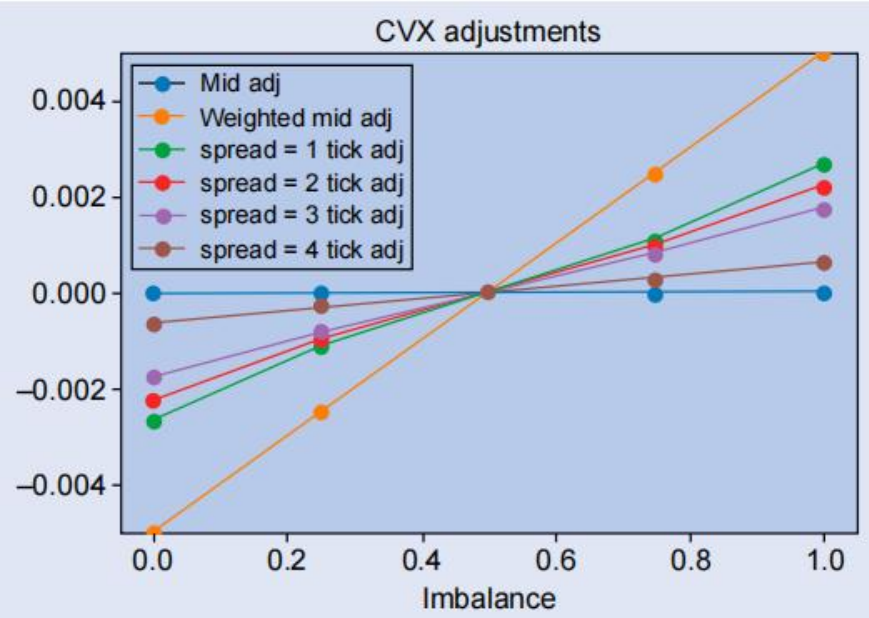
$$G = (I - B)^{-1} * G^1$$

这样得到的G是一个n*m维的列向量，每一个数代表了在该行对应的状态下，Micro-price和Mid-price的差值

计算Micro-price

我们选取了CVX股票2020年第一个交易日估计出的micro price的偏离量进行绘制，得到的结果与论文基本一致。micro price随imbalance的增大而增大。spread越大，对应的偏离量越小，原因在于此时成交的订单不再由于imbalance的不平衡而集中在某一头，而是被分散在中间的价格里，从而由imbalance造成的偏移被削弱了。micro price是介于weighted mid price和mid price中间的价格估计量。





统计量检验

- CJ统计量：$\widehat{\mathrm{CJ}} \equiv N_s/N_r$

$$\Pr(I_t = 1) = \pi = \Phi(\frac{\mu}{\sigma})$$
$$\Pr(Y_t = 1) = \pi_s = \pi^2 + (1-\pi)^2$$

$$CJ = \frac{\pi^2 + (1-\pi)^2}{2\pi(1-\pi)} \geq 1$$

- 游程方法：N_{runs}代表游程总数

$$z = \frac{N_{\mathrm{runs}} - 2n\pi(1-\pi)}{2\sqrt{n\pi(1-\pi)[1-3\pi(1-\pi)]}} \sim N(0,1)$$

$$\pi = \Phi(\mu/\sigma)$$

对应最强的RW1假设

$$P_t = \mu + P_{t-1} + \epsilon_t$$

残差项满足独立同分布（但不一定是正态分布）

统计量检验

❑方差比统计量：

定义方差比为：

$$VR(q) = \frac{Var[r_t(q)]}{\sum_{k=0}^{q-1} Var[r_{t-k}]}$$

在二阶平稳性条件下，方差比是相关系数的线性和：

$$VR(q) = \frac{Var[r_t(q)]}{qVar[r_t]} = 1 + 2\sum_{k=1}^{q-1}(1 - \frac{k}{q})\rho(k)$$

$$\psi^*(q) = \frac{\sqrt{nq}\,(\overline{\mathrm{VR}}(q) - 1)}{\sqrt{\hat{\theta}}} \overset{a}{\sim} \mathcal{N}(0,1) \qquad (2.4.44)$$

对应最弱的RW3假设

$$P_t = \mu + P_{t-1} + \epsilon_t$$

残差项满足两两之间的相关系数为0

1 理论背景

2 价格估计量是否随机

3 价格估计量能否预测交易价格

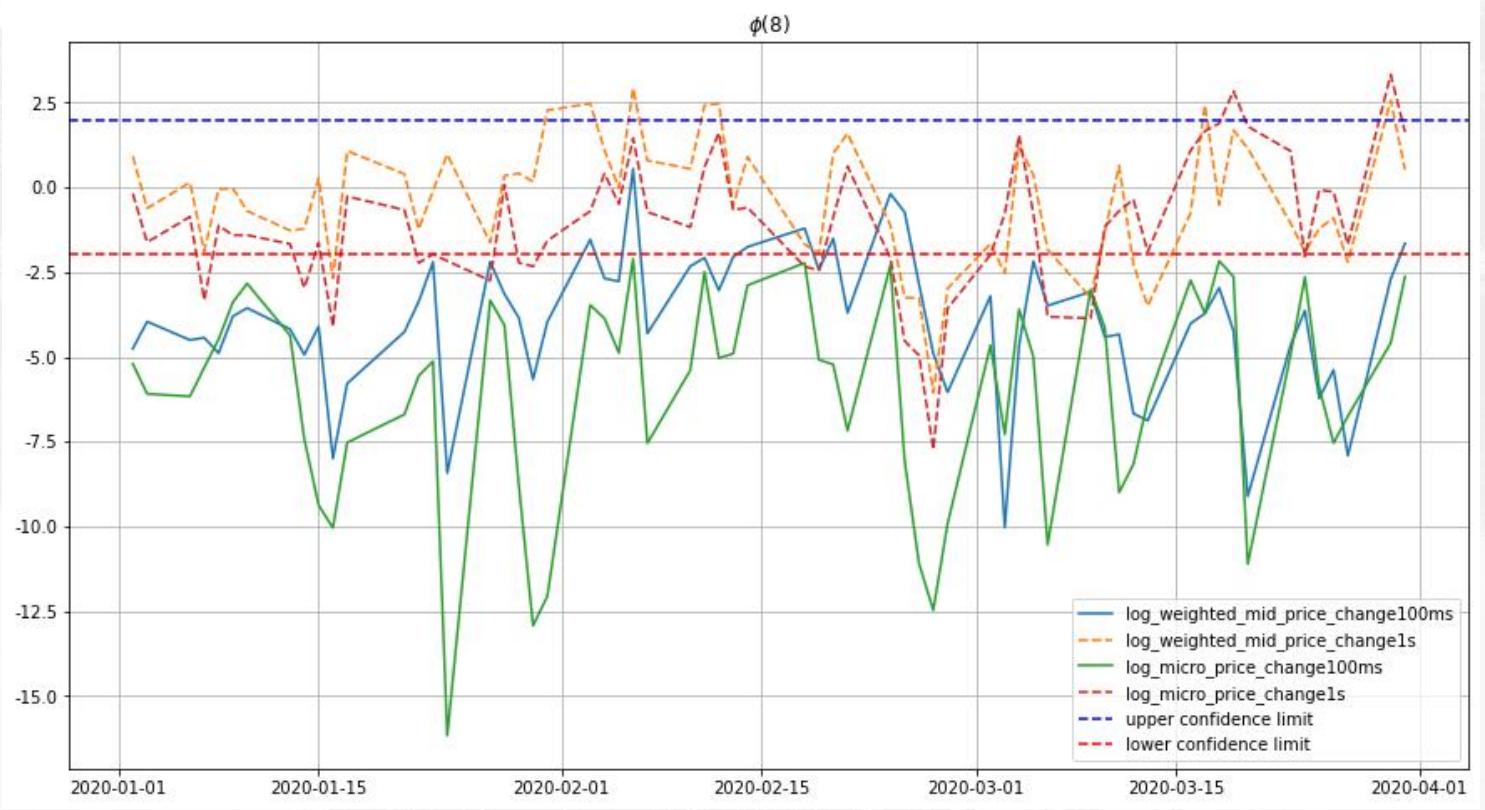
4 总结与展望

个股统计量结果展示

- 慧与科技（Hewlett Packard Enterprise，缩写为HPE）是一家跨国资讯科技公司，总部位于美国加州帕罗奥图，2015年从惠普公司拆分成立，主要为大中型企业提供电脑硬件制造与软件服务。选取这只股票作为代表的原因是它对于高频统计量与低频统计量的分层能力较强

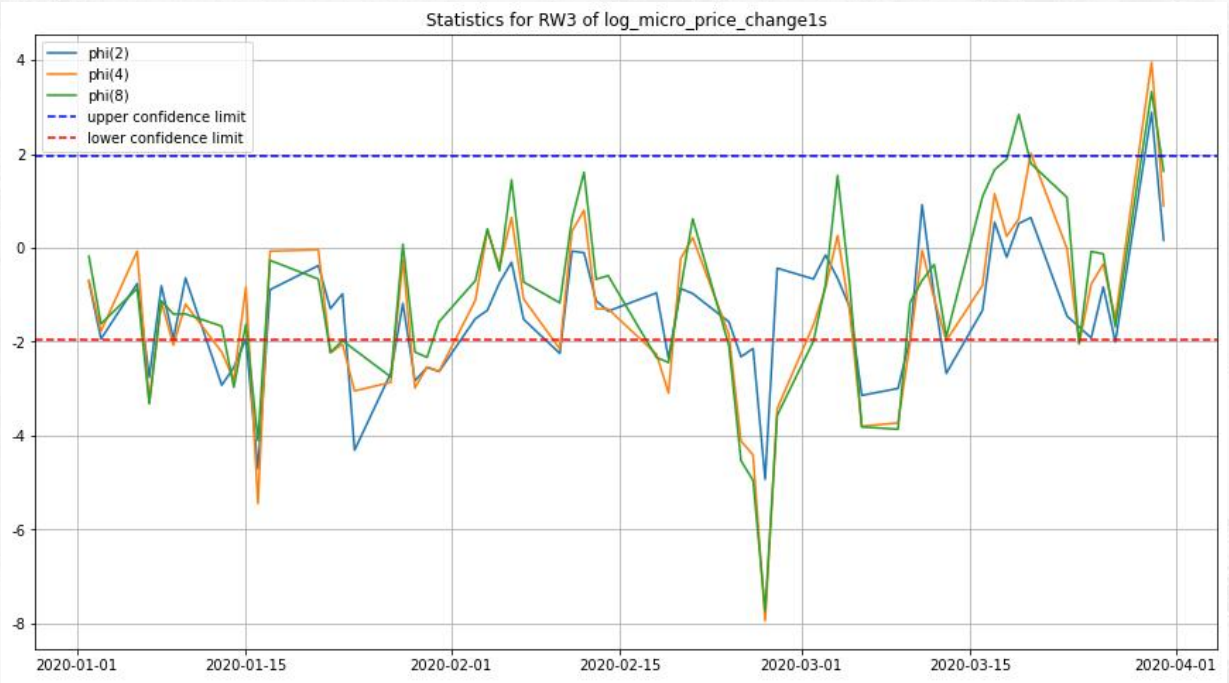
对于这只股票，从虚线相比实线更多地落在置信区间内可以看出，低频数据更加符合RW3假设

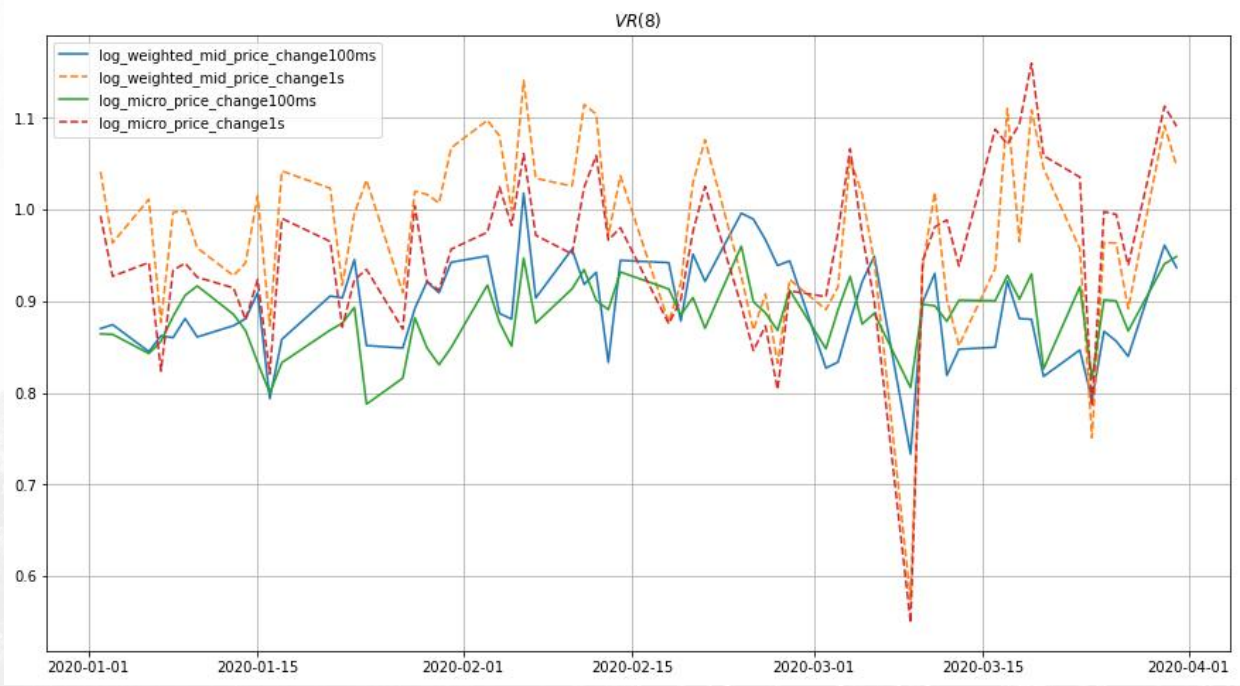
从总体上橙线高于红线，蓝线高于绿线这两个事实可以看出wm相比micro更容易落在置信区间内，从而对这只股票而言wm更加符合RW3假设



个股统计量结果展示

从参数不同的3个方差比统计量的时序图来看，3条曲线走势比较接近，因此在展示对数价格序列随机性的结果时为了方便可以选取其中的一条作为展示对象



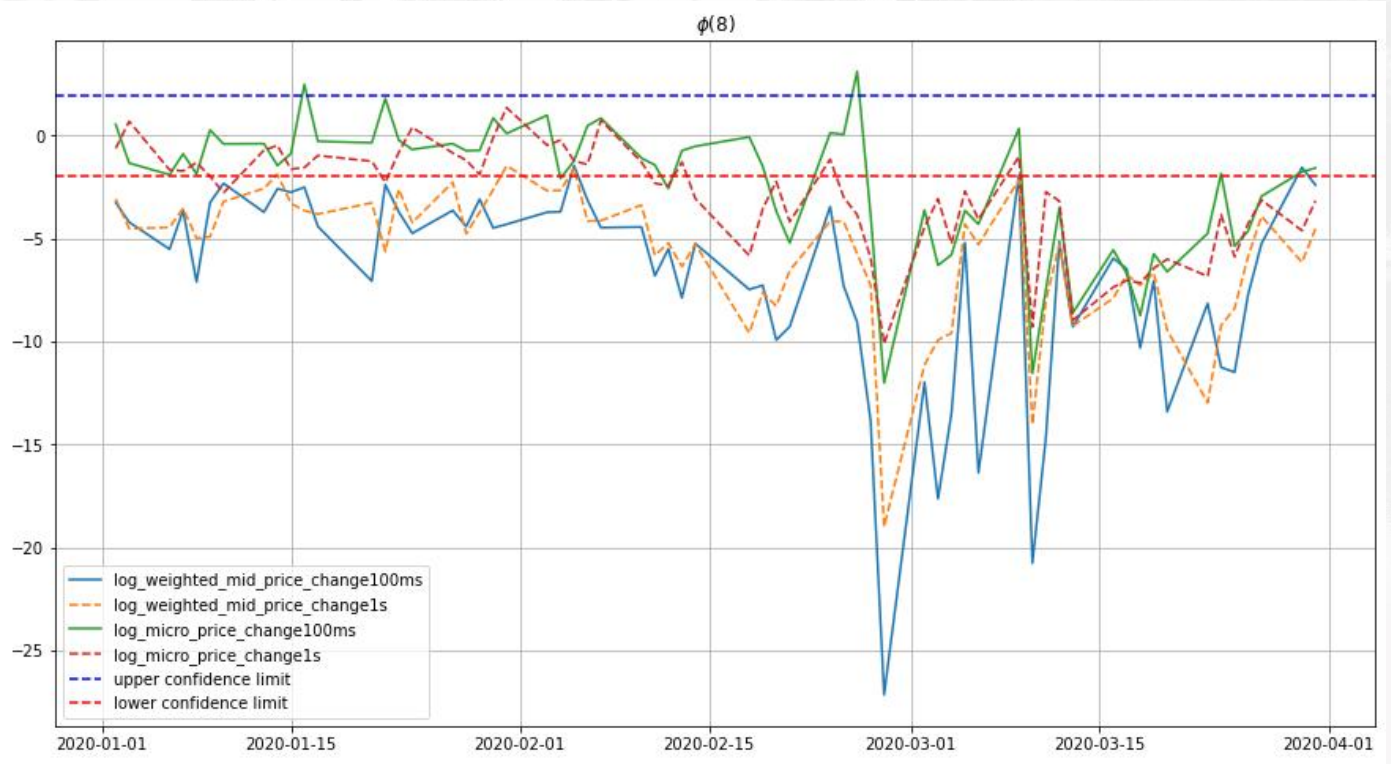


从方差比的时序图来看，大部分方差比位于1以下，说明对数价格差分序列更加容易表现出负的自相关性，其中高频数据处于更低的位置，对应的负相关性更强

个股统计量结果展示

- LyondellBasell Industries NV是一家在荷兰注册成立的跨国化工公司，美国业务总部位于得克萨斯州休斯敦，并在英国伦敦设有办事处。该公司是最大的聚乙烯和聚丙烯技术许可方。它还生产乙烯、丙烯、聚烯烃和含氧燃料。选取这只股票作为代表的原因是它对于不同价格统计量的分层能力较强

从总体上橙线低于红线，蓝线低于绿线这两个事实可以看出micro price相比weighted mid price更容易落在置信区间内，从而对这只股票而言micro price更加符合RW3假设



整体结果展示

- 分别对486支美股和200支A股计算了一季度内两种频率（1s，100ms）下三种价格数据，并计算了对数价格差分序列的统计量，比较三种价格序列的随机性优劣。
- 对于三种价格序列，CJ统计量和z统计量均显著的拒绝了RW1假设。
- 方差比检验则有选择的接受或者拒绝RW3假设，对于一支股票，我们称某种价格优于另一种价格，当且仅当在这一季度内，这种价格方差比检验接受RW3假设的天数超过另一种价格方差比检验接受RW3的天数，即在随机性上此种价格表现得更加好。
- 在整体表现上，Micro-price优于另外两种价格，美股市场上Mid-price优于Weighted mid-price，而在A股市场上Weighted mid-price要优于Mid-price一些。

整体结果展示

- 486支美股，1s频率下，三个方差比统计量中Micro-price不劣于Mid-price的股数是454、462、469，优于Mid-price的股数是244、234、247；不劣于Weighted mid-price的股数是406、399、405，优于Weighted mid-price的股数为312、303、297；100ms频率下，三个方差比统计量中Micro-price不劣于Mid-price的股数是404、422、428，优于Mid-price的股数是221、224、214；不劣于Weighted mid-price的股数是402、414、417，优于Weighted mid-price的股数为309、311、305。
- 200支A股，1s频率下，三个方差比统计量中Micro-price不劣于Mid-price的股数是197、199、200，优于Mid-price的股数是181、176、183；不劣于Weighted mid-price的股数是148、150、136，优于Weighted mid-price的股数为124、120、110；100ms频率下，三个方差比统计量中Micro-price不劣于Mid-price的股数是191、190、192，优于Mid-price的股数是173、168、172；不劣于Weighted mid-price的股数是178、175、171，优于Weighted mid-price的股数为157、153 、147。

1 理论背景

2 价格估计量是否随机

3 价格估计量能否预测交易价格

4 总结与展望

价格与过去价格估计值的相关性

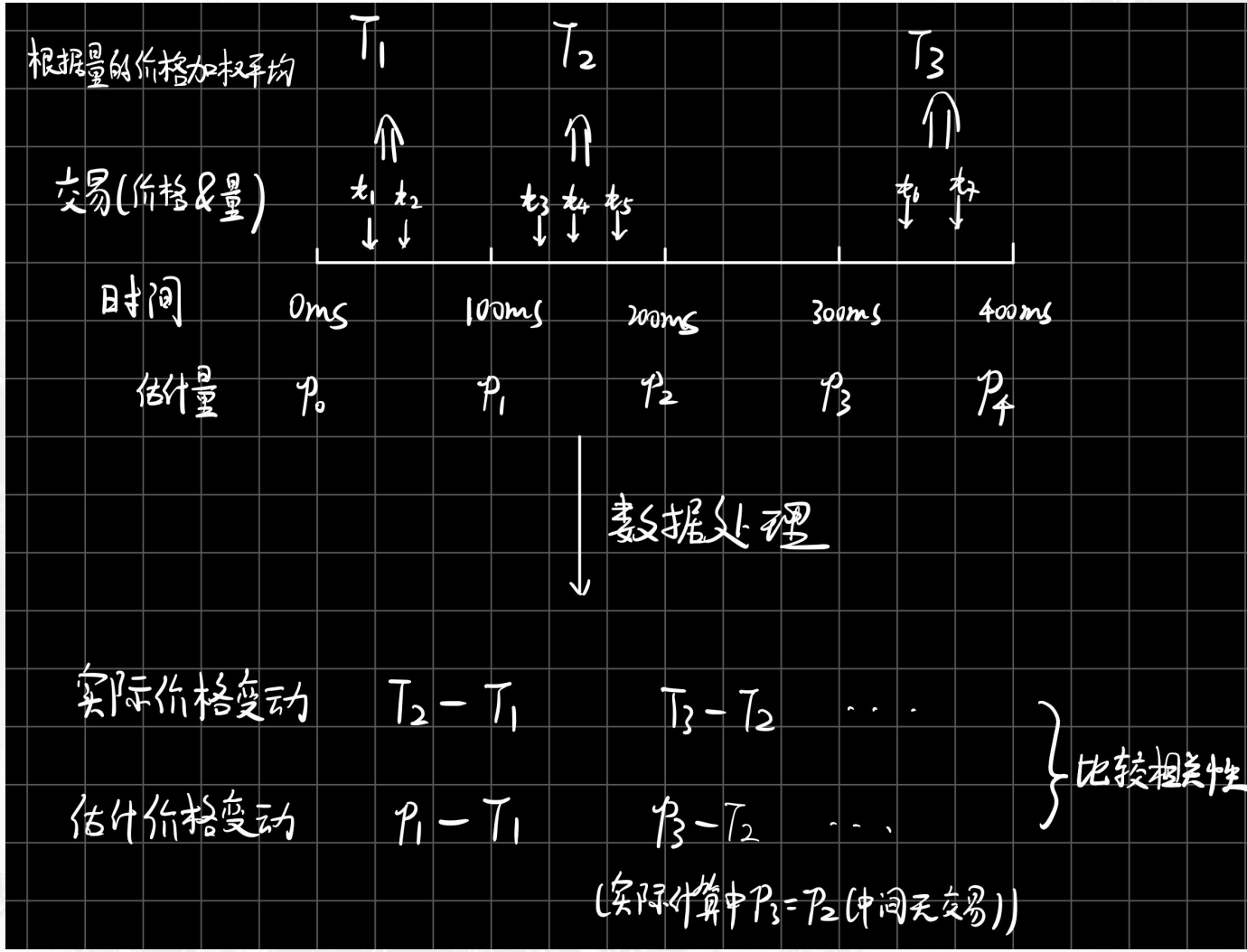
- 我们知道，一个良好的价格应该满足随机游动假设。那么，根据过去的价格，我们应该很难预测商品的当前价格。否则，一旦商品的价格可预测，那么就可以认为可能市场存在套利机会。

- 因此，我们希望考察股票的交易价格与上一时刻的价格估计量的相关系数。如果某个估计量存在相对较小的相关性，那么可以一定程度上说明这个估计量要“优于”其他估计量。

- 为此，我们对第一季度美股数据的100支股票的交易价格与前一时刻（比如，如果交易时间发生在某分钟的第t/10秒，那么就考虑其第[t]/10秒时的价格估计值，因为我们当前的价格统计精度只能到100ms）的价格估计量进行研究。

价格与过去价格估计值的相关性

- 我们决定采取以下检验方案：
- 将交易价格按照每100ms打包做加权平均，得到新的每100ms的打包价。将相邻两次打包价的价差作为实际价格变化差，将每100ms的估计价格减去前一次打包价作为估计价格变化差，比较二者的相关系数作为评判价格预测能力的标准
- 实际数据中很多时候存在100ms间没有交易发生，在计算过程中我们就舍去这段时间（事实上，一段没有交易发生的时间前后价格估计量也是一致的，这保证了舍去时间的合理性）



价格与过去价格估计值的相关性

- 最终，我们得到了美股100只股票的相关系数结果，如下图展现了100只股票的结果的一些统计量（pearson，spearman，kendall是三种不同的计算相关性的方法）

	mid with pearson	wm with pearson	micro with pearson	mid with spearman	wm with spearman	micro with spearman	mid with kendall	wm with kendall	micro with kendall
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.159785	0.144415	0.159177	0.178035	0.166368	0.168145	0.132136	0.122061	0.121671
std	0.114898	0.112904	0.114561	0.113158	0.111322	0.109041	0.088016	0.084340	0.081097
min	-0.219298	-0.213670	-0.219258	-0.021911	-0.033637	-0.005643	-0.012368	-0.019996	-0.004228
25%	0.084874	0.083600	0.086227	0.124817	0.103720	0.099304	0.090384	0.075546	0.072058
50%	0.157807	0.140096	0.157397	0.169588	0.153026	0.165383	0.122623	0.109801	0.118036
75%	0.204001	0.179253	0.203127	0.223741	0.197661	0.214463	0.167341	0.145911	0.155106
max	0.703574	0.703151	0.703593	0.436236	0.447595	0.435624	0.339623	0.343198	0.323628

- 可以发现，mid price的相关性确实更大，而wm表现的更好，这和论文的结果是一致的。

价格与过去价格估计值的相关性

- 除此之外，我们还有以下结果：
- 在前100支股票中，三种方法中相关系数mid>micro>wm的股票个数为(51, 63, 64)
- 在前100支股票中，三种方法中相关系数mid>wm>mirco的股票个数为(9, 10, 14)，说明mid相关性最大的股票占了较大比例，符合我们的预期。
- 更直观地，在前100支股票中，三种方法中相关系数mid，micro，wm相关系数最大（估计量较差）与最小（估计量较好）的股票个数如下图：

最大个数：

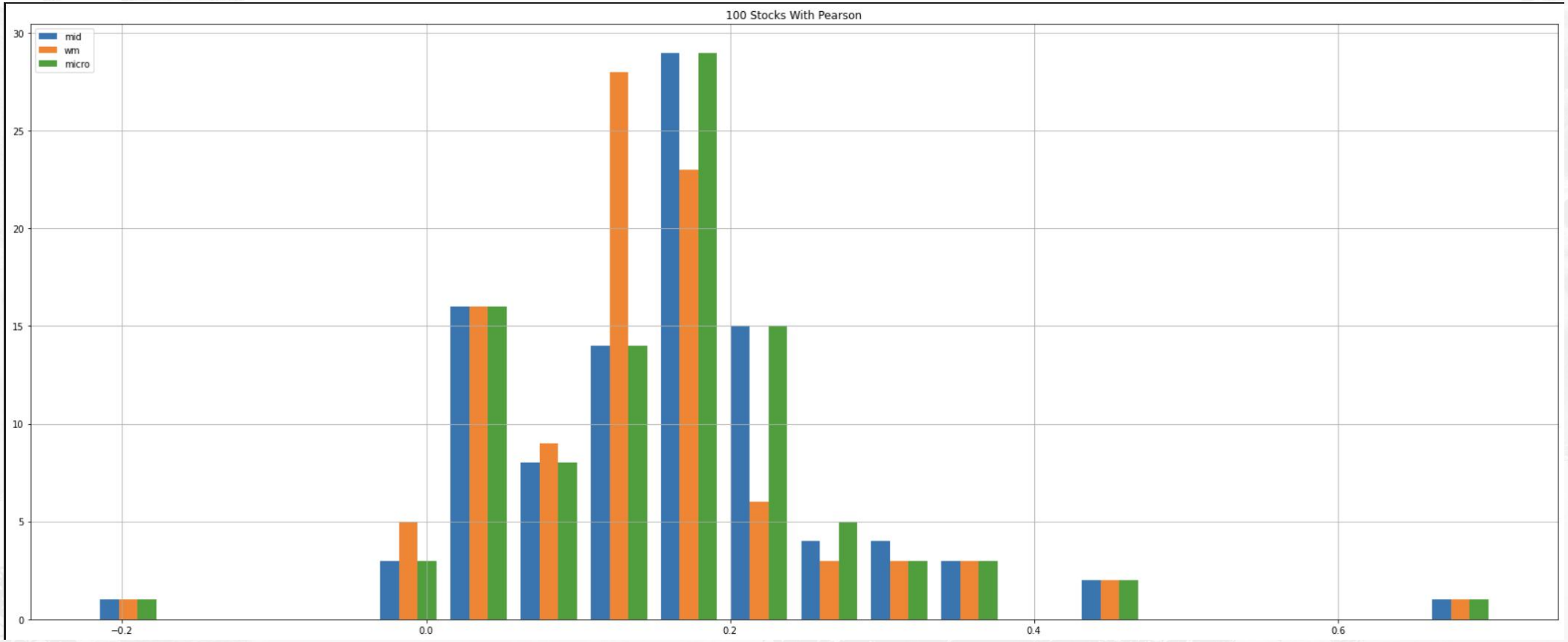
	mid	wm	micro
pearson	60	9	31
spearman	73	16	11
kendall	78	14	8

最小个数：

	mid	wm	micro
pearson	8	80	12
spearman	10	71	19
kendall	9	69	22

价格与过去价格估计值的相关性

- 作为补充，我们展示100支股票相关系数的频数分布直方图，纵轴代表股票数量。从图中可以明显看出由于weighted mid price在0.1附近的相关系数较多，导致了weighted mid price相关系数的中位数和平均数都比较小，进而得到weighted mid price 是较优秀的价格估计量的结论



1 理论背景

2 价格估计量是否随机

3 价格估计量能否预测交易价格

4 总结与展望

总结与展望

- 我们对A股市场的200支历史最悠久的股票和标普500的486支成分股分别检验了三种价格估计量是否满足随机游动假设，比较了它们的差异，认为总体上micro price是最接近RW3假设的，但是不同股票间差异较大。探索了采样频率的影响。
- 我们对标普500中的100支成分股探索了价格估计量能否预测实际交易价格，通过计算估计量与实际价格的相关系数得出weighted mid price的预测能力最差，是三种估计量中最好的。
- 未来我们会利用并行计算对A股全部近2000支股票和标普500的所有成分股进行上述研究。

北京大学
PEKING UNIVERSITY

谢谢！