高频因子的优势：与低频因子相比，高频数据在量化选股中的优势主要体现在：因子拥挤度相对较低、因子多样性好、检验因子的独立样本多。

研究内容：本报告从四类不同的角度构建因子：日内价格相关因子、日内价量相关因子、盘前信息因子、特定时段采样因子。考察了55个因子周频选股的表现。其中，日内价格相关的因子是由日内收益率的高阶统计量和日内价格形态衍生的因子；日内价量相关因子包括成交量分布以及用价量关系构建的因子；盘前信息因子主要是从开盘集合竞价信息中提炼的因子；特定时段采样因子主要是指根据一定规则筛选出重要时段，在该时段采样提取的因子。

实证分析：采用因子IC和多空收益进行分析，筛选出12个周度选股能力较好的因子：real_skew（已实现偏度）、ret_intraday（日内收益率）、ratio_volumeH8（尾盘半小时成交量占比）、corr_VP（日内价量相关性）、corr_VRlag（量与滞后收益率相关性）、Amihud_illiq（Amihud非流动性因子）、ret_open2AH1（开盘价相对第一阶段集合竞价最高价的收益率）、ret_open2AL1（开盘价相对第一阶段集合竞价最低价的收益率）、ret_H8（尾盘半小时收益率）、real_skewlarge（大成交量已实现偏度）、corr_VPlarge（大成交量）、corr_VRlaglarge（大成交量量与滞后收益率相关性）。

从因子表现看，本报告筛选出来的因子都展示出了不错的多空超额能力。

一、高频因子思考：从低频信息到高频信息

近年来，A股市场机构化趋势明显，量化私募机构的惯例规模也迅速扩大，产生了一批惯例规模超过百亿的量化私募机构。与此同时，传统的风格因子波动增大，从市场获取超额收益的难度在增加。

因子拥挤是因子收益下降的原因之一。因子代表着市场某方面的非有效性、或者是一段时期内的定价失效。当某类因子收益高的时候，会吸引更多的资金进入，从而出现因子拥挤，降低因子的预期收益。一旦新的因子被公开，套利资金的介入会使得错误定价收窄，因子收益也会跟着下降。因此，在多因子选股模型中，因子的开发和更新迭代变得越来越重要。

以传统日频价量和更低频财务数据为基础的因子开发是一种研究途径。由于基础因子广为人知，在此基础上进行因子挖掘的收益提升空间相对有限。而且日频数据由于本身的数据量和信息量有限，过度挖掘会增大过拟合的风险。

以高频价量数据为基础的因子开发在当下具有更大的收益提升空间。与低频因子相比，高频数据在用于量化投资中存在一定优势。

首先，高频价量数据的体量明显大于低频数据。以分钟行情为例，用压缩效果较好的mat格式存储2020年全市场股票的分钟行情数据（包括分钟频的高开低收价格数据、买卖盘挂单数据等），约为12GB。如果是快照行情（目前上交所和深交所都是3秒1笔）或者level2行情，数据量要大很多。因此，高频数据因子挖掘对信息处理能力和处理效率的要求较高。而且，日内数据，尤其是level2数据，一般要额外付费，甚至需要自行下载存储实时行情，在此基础上构建的因子拥挤度较低。

其次，高频价量数据一般是多维的时间序列数据，数据中噪声比例较高，而且与ROE、PE这类低频指标本身就具有选股能力不同的是，原始的高频行情数据一般不能直接用作选股因子，而要通过信号转换、时间序列分析、机器学习等方法从高频数据中构建特征，才能作为选股因子。此类因子与低频信号的相关性较低，而且由于因子开发流程相对复杂，不同投资者构建的因子更具有多样性。

此外，高频数据开发的因子一般调仓周期较短，意味着在检验因子有效性的时候，同一段测试期具有更多的独立样本。例如，在一年的测试期内，只有12个独立的样本段用于检验月频调仓的因子，与之相比，有约50个独立的时段用于检验周频调仓因子，有超过240个独立的时段用于检验日频调仓的因子。独立样本呢的增多有助于检验高频银子的有效性。

高频数据挖掘因子的难点在于数据维度大、噪声高。凭借专业投资者的经验或者是参阅已发表的文献，可以从高频数据中提炼出一部分有选股能力的特征。此外，机器学习方法擅长从数据中寻找规律和特征，是高频数据因子挖掘的有力工具。本篇专题报告通过将日内的数据分成几个数据维度，从不同的数据维度中建模提取选股因子。

二、因子构建方法和主要性能指标

（一）因子构建方法

本报告构建因子时，从四大类不同的角度构建因子：日内价格相关因子、日内价量相关因子、盘前信息因子、特定时段采样因子。其中，日内价格相关的因子是由日内收益率的高阶统计量和日内价格形态衍生的因子，共计10个因子。日内价量相关因子包括成交量分布以及价量关系构建的因子，共计13个因子。盘前信息因子主要是从开盘集合竞价信息中提炼的因子，共计7个因子。特定时段采样因子主要是指根据一定规则筛选出重要时段，在该时段采样并计算提取的因子，主要包括尾盘数据构建的因子和大成交量时段构建的因子，共计25个因子。

（二）因子分析指标

收益预测时，统一按照周度调仓的假设进行分析。T日收盘后计算的因子预测的是从T+1日股票未来5个交易日的收益率，因子IC测试和收益回测都按照T+1日至T+5日进行计算。

股票池：全市场或者中证500指数成分股，剔除新股、ST个股、因子计算当日和交易日停牌或者涨跌停的个股。

IC：T日因子与T+1日开盘后5个交易日收益率的相关系数。

IC胜率：IC与因子方向相同的比率（因子方向按照整个回测区间因子IC均值的正负号确定）。

年化ICIR：IC绝对值与IC标准差之比的年化值，即

ICIR=sqrt(N)*abs(IC)/std(IC)

其中N表示一年内包含的交易周数。

在选股测试时，按照因子方向和因子取值将股票池内股票平均分为10组组内股票等权。

多空超额收益率：因子多头组合收益相对因子空头组合的年化超额，越大则表示因子选股能力越强。

多空超额胜率：因子多头组合收益相对因子空头组合的超额胜率。

正Alpha：因子多头组合收益相对基准指数的年化超额，越大则表示因子多头相对基准的超额收益越高。

负Alpha：因子空头组合收益相对基准指数收益的年化超额，越小（为负且绝对值越大）则表示因子空头相对基准负的超额收益越高。

本报告的多空超额收益率、正Alpha、负Alpha是分别根据各自组合的累计超额计算出来的，由于存在复利效应，因子的正Alpha、负Alpha之差一般与因子多空超额收益率略有差异。

三、日内价格相关因子

（一）因子计算方法

价格数据中蕴含了丰富的股票信息，在传统多因子体系中，股价反转、波动率都是重要的风格因子。在高频数据中，可以用类似方法构建相关的特征。本报告根据收益率的高阶统计量构建了如下选股因子：

表1：日内收益率高阶统计量因子

因子名	因子描述
real_var	已实现方差，分钟行情收益率的方差
real_skew	已实现偏度，分钟行情收益率的偏度
real_kurtosis	已实现峰度，分钟行情收益率的峰度
real_upvar	上行收益率方差，仅考虑收益率大于0时的分钟行情收益率方差
real_downvar	下行收益率方差，仅考虑收益率小于0时的分钟行情收益率方差
ratio_realupvar	上行收益率方差比值，real_upvar/real_var
ratio_realdownvar	下行收益率方差比值，ratio_realdownvar/real_var

$$r_{t,D,i} = \frac{p_{t,D,i}}{p_{t-1,D,i}} - 1, i = 2, 3, \ldots, T$$

$$real\_var_{D,i} = \frac{1}{T-2}\sum_{t=2}^{T}(r_{t,D,i} - \overline{r}_{D,i})^2$$

$$real\_skew_{D,i} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{(r_{t,D,i} - \overline{r}_{D,i})^3}{real\_var_{D,i}{}^2}$$

$$real\_kurtosis_{D,i} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{(r_{r,D,i} - \overline{r}_{D,i})^4}{real\_var_{D,i}{}^2}$$

$$ratio\_realupvar_{D,i} = \frac{real\_upvar_{D,i}}{real\_var_{D,i}}$$

$$ratio\_realdownvar_{D,i} = \frac{real\_downvar_{D,i}}{real\_var_{D,i}}$$

根据日内股价的形态，可以构建如下因子。

表2：价格形态衍生因子

因子名	因子描述
trendratio	趋势占比，日内价格变化/分钟频价格变化绝对值之和
ret_intraday	日内收益率，收盘价/开盘价-1
intraday_maxdrawdown	日内最大回撤，日内分钟频行情的最大回撤

$$trendratio_{D,i} = \frac{p_{T,D,i} - p_{1,D,i}}{\sum_{t=2}^{T} |p_{t,D,i} - p_{t-1,D,i}|}$$

$$ret\_intraday_{D,i} = \frac{P_{T,D,i}}{open_{D,i}} - 1$$

$$intraday_maxdrawdown_{D,i} = ?$$

（二）因子全市场选股、中证500内表现

从下表中可以看出，日内价格数据构建的因子中，real_var、real_upvar、intraday_maxdrawdown因子在全市场及中证500中的效果较好，且最近几年这些因子整体上维持着较为有效的特征。

四、日内价量相关因子

（一）因子计算方法

成交量是日内行情信息的重要组成部分。一方面，成交量的分布可以反映投资者的行为特征，另一方面，成交量与价格或者价格走势的关系可以确认价格形态的信息。

将每天的4小时交易时间按照时间平均分为8段，根据每段的成交量占全天成交量之比，构建如下因子：

表5：成交量分布因子

因子名	因子描述
ratio_volumeH1	早盘成交量占比：开盘后第1个半小时成交量占全天成交量之比
ratio_volumeH2	成交量占比H2：开盘后第2个半小时成交量占全天成交量之比
ratio_volumeH3	成交量占比H3：开盘后第3个半小时成交量占全天成交量之比
ratio_volumeH4	成交量占比H4：开盘后第4个半小时成交量占全天成交量之比
ratio_volumeH5	成交量占比H5：开盘后第5个半小时成交量占全天成交量之比
ratio_volumeH6	成交量占比H6：开盘后第6个半小时成交量占全天成交量之比

因子名	因子描述
ratio_volumeH7	成交量占比H7：开盘后第7个半小时成交量占全天成交量之比
ratio_volumeH8	尾盘成交量占比：开盘后第8个半小时成交量占全天成交量之比

考虑到价格和成交量的相互关系，可以构建以下因子：

表6：价量相关因子

因子名	因子描述
corr_VP	价量相关性，分钟成交量与价格相关性
corr_VR	收益率与量相关性，分钟成交量与收益率相关性
corr_VRlag	量与滞后收益率相关性，分钟成交量与上一时刻收益率相关性
corr_VRlead	量与超前收益率相关性，分钟成交量与下一时刻收益率相关性
Amihud_illiq	Amihud 非流动性因子

其中，价量相关性是指价格序列和成交量序列的相关性，记股票日内分钟频率下的成交量序列为{v_{t,D,i}},t=1,2,3,...,T,则该因子计算方法为

$$corr\_VP_{D,i} = corr(v_{t,D,i}, p_{t,D,i})$$

收益率与量的相关性为

$$corr\_VR_{D,i} = corr(v_{t,D,i}, r_{t,D,i})$$

$$corr\_VRlag_{D,i} = corr(v_{t,D,i}, r_{r-1,D,i})$$

$$corr\_VRlead_{D,i} = corr(v_{t,D,i}, r_{t+1,D,i})$$

Amihud非流动性因子是Amihud在2002年提出了衡量流动性的因子，考虑单位成交额驱动下，股价的变化幅度。因子值越大，说明股票的价格越容易被交易行为所影响（即流动性越低）。常见的Amihud非流动性因子是按照日频构建的，本报告在分钟频率下构建类似的因子。

$$Amihud_illiq_{D,i} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{|r_{t,D,i}|}{p_{t,D,i}v_{t,D,i}}$$

该因子是指在分钟频率下，单位成交额驱动下，股价的变化幅度。Amihud非流动性因子取值非负。

（二）因子全市场选股、中证500选股内表现

日内价量相关因子表现如下表所示。从下表中可以看出，日内价量类数据构建的因子中，ratio_volumeH1、ratio_volumeH5、corr_Vrlag、Amihud_illiq这5个因子在全市场及中证500中的选股效果较为显著。

五、盘前信息因子

（一）因子计算方法

盘前信息主要包括隔夜收益率（开盘价相对前收盘的收益率）和开盘前集合竞价信息。目前，A股证券交易所在每隔交易日的9:15到9:25为开盘集合竞价时间。开盘集合竞价又分为两个阶段，其中第一阶段是9:15至9:20，该阶段允许撤销已经提交的订单；第二阶段是9:20至9:25，该阶段目前不允许撤销已经提交的订单。集合竞价信息反映出资金的试盘行为和多空双方的博弈。本报告考察隔夜收益率和集合竞价的相关因子如下所示。

表9：盘前信息因子列表

因子名	因子描述
ret_overnight	隔夜收益率，开盘价相对前收盘价的收益率
ret_open2AH1	开盘价相对第一阶段集合竞价最高价的收益率
ret_open2AL1	开盘价相对第一阶段集合竞价最低价的收益率
ret_open2AH2	开盘价相对第二阶段集合竞价最高价的收益率
ret_open2AL2	开盘价相对第二阶段集合竞价最低价的收益率
diverge_A1	第一阶段集合竞价振幅
diverge_A3	第二阶段结合竞价振幅

（二）因子全市场选股、中证500选股内表现

从下表中可以看到，盘前相关的信息中，ret_open2AH、diverge_A2因子在全市场及中证1000中的选股效果较为显著。

六、特定时段采样因子

（一）因子计算方法

本报告将部分时段的数据进行重点分析，产生衍生因子。一般来说，开盘后半小时（9点半至10点）和收盘前半小时（14点半至收盘）的股票成交活跃，多空博弈激烈，蕴含的信息相对较多。本报告针对收盘前半小时的价量信息构建了如下因子。

表12：开盘后半小时因子列表

因子名	因子描述
ret_H1	开盘后半小时的收益率
ret_close2H1	开盘后半小时至收盘的收益率
corr_VPH1	开盘后半小时的 corr_VP
corr_VRH1	开盘后半小时的 corr_VR
corr_VrleadH1	开盘后半小时的 corr_VRlead
corr_VrlagH1	开盘后半小时的 corr_VRlag
real_varH1	开盘后半小时的 real_var
real_kurtosisH1	开盘后半小时的 real_kurtosis

因子名	因子描述
real_skewH1	开盘后半小时的 real_skew

表13：收盘前半小时因子列表

因子名	银子描述
ret_H8	收盘前半小时的收益率
corr_VPH8	收盘前半小时的 corr_VP
corr_VRH8	收盘前半小时的 corr_VR
corr_VRleadH8	收盘前半小时的 corr_VRlead
corr_VRlagH8	收盘前半小时的 corr_VRlag
real_varH8	收盘前半小时的 real_var
real_kurtosisH8	收盘前半小时的 real_kurtosis
real_skewH8	收盘前半小时的 real_skew

在不同的成交中，大单成交与主力资金关联较大，蕴含的信息可能更多。本报告将个股在每隔交易日的分钟成交量时间序列按照成交量大小排序，将分钟成交量排名前1/3的成交量定义为“大成交量”。针对大成交量对应的时刻的股价信息，可以构建大成交量相关因子。

表14：大成交量相关因子列表

因子名	因子描述
real_varlarge	大成交量对应的收益率方差
real_kurtosislarge	大成交量对应的收益率峰度
real_skewlarge	大成交量对应的收益率偏度
ratio_realvarlarge	大成交量方差占比，real_varlarge/real_var
corr_VPlarge	大成交量对应的 corr_VP
corr_VRlarge	大成交量对应的 corr_VR
corr_VRleadlarge	大成交量对应的 corr_VRlead
corr_VRlaglarge	大成交量对应的 corr_VRlag

（二）因子全市场选股、中证500选股内表现

开盘后半小时构建的因子中，其中ret_close2H1、real_varH1在全市场及中证500中的选股效果较为显著。

收盘前半小时构建的因子中，其中ret_H8、corr_VRH8在全市场及中证500中的选股效果较为显著。

大成交量对应的时点构建的因子中，其中real_varlarge、ratio_realvarlarge、corr_Vplarge因子在全市场及中证500中的选股能力较为显著。

七、总结和展望

本报告从四类不同的角度构建因子：日内价格相关因子、日内价量相关因子、盘前信息因子、特定时段采样因子。考察了55个因子周频选股的表现。

采用因子IC和多空收益进行分析、筛选出12个周度选股能力较好的因子：real_skew（已实现偏度）、ret_intraday（日内收益率）、ratio_volumeH8（尾盘半小时成交量占比）、corr_VP（日内价量相关性）、corr_VRlag（量与滞后收益率相关性）、Amihud_illiq（Amihud非流动性因子）、ret_open2AH1（开盘价相对第一阶段集合竞价最高价的收益率）、ret_open2AL1（开盘价相对第一阶段集合竞价最低价的收益率）、ret_H8（尾盘半小时收益率）、real_skewlarge（大成交量已实现偏度）、corr_VPlarge（大成交量价量相关性）、corr_VRlaglarge（大成交量量与滞后收益率相关性）。

从因子表现来看，本报告筛选出来的因子都展示除了不错的多空超额能力。

如何更好的将上述高频因子的选股能力转化为多头的超额收益，是非常有实际意义的课题。