流动性二期：预测

项目目标

我们希望基于过去的债券特征，根据表现相似的债券，得到对于目标债券流动性指标未来的可能分布情况，并给出预测

债券特征的提取：
1.股价情况
2.交易情况
3.基本面情况（我们认为基本面情况本身反映在了具体的交易以及股价信息中，故只考虑了剩余期限以及中债估）

流动性指标：
1.下一笔交易间隔多少天**
2.未来一段时间（7天）内交易笔数
3.未来一段时间（7天）内交易笔数的变化
4.未来一段时间（7天）内有交易的天数
5.未来一段时间（7天）内的总交易量

- 在先期研究中发现2-5指标其实与标注数据有较好的相关性，故将作为我们最关心的指标

SQL取数

bid&ofr数据

- 以下功能均基于bid&ofr.py实现
- 取数按照交易日取了bid以及ofr的每笔价格（收益率，净价口径待有）
- 最终每天对于每只债券计算收益率/净价的均值、中位数、75%分位数、最大值、最小值以及相应笔数，最终的merged_statistics.csv文件即是以上每天计算的指标合起来的结果

交易数据

- 以下功能均基于deal.py实现
- 取数时分为两步取，在cfets相关数据库中取每个交易日的总量（volume），在broker相关数据库中取逐笔的交易信息，取收益率时选择了ytm作为收益率避免负值
- 最终每天对于每只债券计算收益率/净价的均值、中位数、75%分位数、最大值、最小值以及相应笔数和总量，最终的merged_deal_statistics.csv文件即是以上每天计算的指标合起来的结果

中债估数据

- 以下功能均基于cbv_updated.csv实现
- 具体取到期收益率以及剩余期限（以*为单位，数值类型为float）
- 中债估需要保证其将日期向后移动一个交易日，即今天的价格对应昨天的中债估
- 除此以外，中债估数据可能在同一个issue_date以及bond_uni_code下有多个不同的值，此时选择取剩余期限最小的
- 最终的merged_cbv_statistics.csv文件即是以上每天取完指标合起来的结果，包含每天每个券的剩余期限、到期收益率

指标计算

1.合并相关文件

- 需要合并的文件包括bond_type.csv（后续筛选做信用/利率债时使用）以及上述的merged_XXX.csv文件
- 合并时会再针对不在trading_date.csv中的日期进行筛选，保证只有交易日

2.计算天数有关指标

- 由于我们后续需要计算距离上一个/下一个交易日多少天这个指标，故我们需要针对日期进行填充
- 为了方便我们的计算，我们对于每一行是否有bid/ofr/deal进行标记，并记录有以上记录的行的行索引，上一个有以上记录的行的行索引
- 我们计算的天数相关指标包括：
1.距离上一个/下一个交易日（不含今天）的天数
2.在过去/未来N天（不含今天）交易的天数
3.在未来/过去内交易的天数（不含今天）

3.时间窗口指标的计算

- 变化指标包括：
1.中债估变化
2.bid/ofr/deal的yield变化
3.bid/ofr/deal交易笔数的变化
- 总和指标：
1.bid/ofr/deal在时间窗口内的笔数总和
2.deal volume在时间窗口内的总和
- 滑动平均指标：ofr/bid/deal的价格加权平均，平均方式也为刚刚加权
  - 给定一个的交易日，我们选择过去7/14天（不含当天）为过去的时间窗口，选择未来7天为未来时间窗口，分别对以上三个时间窗口计算以上指标
  - 1.基于行索引的差计算距离上一个/下一个交易日的天数
  - 2.基于rolling函数计算一段时间内的总交易天数
  - 变化指标基于对给定时间窗口内的变化量进行加权平均实现。
  - 具体加权方式为底数为0.98的指数加权，距离当前日期越近权重越大
  - 实现方式仍为rolling函数

4.ATR

- 衡量波动性的一个指标，常用于股票交易
- 计算方式：max(当天最高收益率，当天最低收益率，|当天最高 上一天平均，|当天最低 上一天平均|)
  - 对bid/ofr数据在过去的时间窗口下计算这个值的滑动平均，方法同之前

对于数据的一些补充说明

最终的每个bond_uni_code，每个issue_date对应的指标情况如下

- 股价端6交易端：
1.时间窗口内总交易笔数，总交易天数
2.距离上一个/下一个交易日相差的总日天数
3.时间窗口内收益率/收益率变化/笔数变化的加权平均
4.时间窗口内ATR的加权平均
- 基本面：
1.时间窗口内中债估收益率变化的加权平均
2.剩余期限
3.中债总收益率

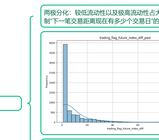
利率债的券总数：3772

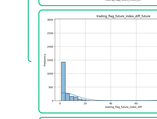
由于我们采取了日期填充，故每一只券一定对应固定的区间长度（即交易日天数），数据的总行数在27w左右

- 以上数据依旧不能直接作为我们需要操作做的数据集，因为其中很大一部分数据都呈现NAN，我们需要进一步筛选以保证可操作性
- 在以“距离过去交易天数”（这个筛选指标起到了较大的作用，实际上选择了我们要聚类的所有指标）不含NAN这个条件筛选后，我们大致余下了4.54只券，1.8w条左右的数据

数据分布的一些问题

- 两极分化：较低流动性以及极高流动性占大头（这里验制“下一笔交易距离现在有多少个交易日”的分布情况）





数据缺失值的问题

- 缺失的指标为我们关注的“下一笔交易距离现在多少个交易日”
- 产生缺失的原因：在可以找到的未来里，没有下一笔交易，即相当低的流动性水平
- 上图中我们对所有这样的缺失值填充了99

基于指标进行聚类

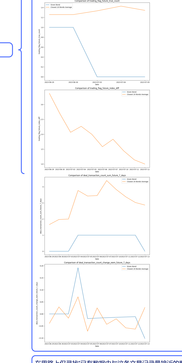
1.指标选择

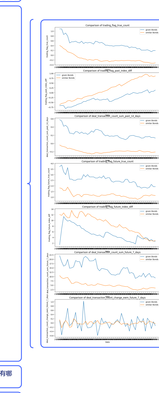
- 报价端：
1.bid/ofr在过去14天内的总笔数
2.bid/ofr距离上一个交易日多少天
3.14天内bid/ofr的总天数
- 交易端：
1.14天内交易的总笔数
2.14天内交易的总量
3.距离上一次交易的交易日数
4.14天内交易的总天数
- 基本面：
1.剩余期限
2.中债估收益率
  - 补充说明：指标选择的多会存在边际改善低。特征的提取分析效率更低的问题。更高维度的马氏距离计算也存在计算资源问题

2.基于马氏距离进行聚类操作

1.券与券之间的聚类操作

- 在思路上试图寻找与“相应券在给定时间段表现相似的债券”，实现方式上基于寻找在这一时间段内距离特定券马氏距离最近的N只债券实现
- 问题在于：
1.更多债券的引入带来了方差的下降，难以观察到给定债券和临近债券的流动性指标同步变化
2.结果特异性较强（给定了一只具体债券）
3.可用数据有限，如果以券为单位没有约400券不同数据可用
  - 尝试多时多关系后呈现了更好的同步变化趋势，一定程度说明了基于聚类思想预测的可行性





2.券与券之间的聚类操作

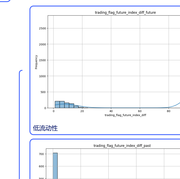
- 在思路上仅寻找“已有数据中与这条交易记录最接近的数据有哪些”，基于它们的未来流动性指标构建分布，进行预测
- 在该思路下，仅能够保证某一日的交易数据对应过去某一日非同一个券的交易数据，日期方面没有进行任何筛选
- 具体的操作方式有以下两种：
1.通过选择合适的分布进行记忆，以此给出一个连续的分布函数，我们可以直到某一天发生交易的可能性为多少
2.不关注具体的分布形式，而是采用离散的区间处理，只知道特定的时间区间，给出“下一笔交易发生在特定的时间区间”概率是多少

评价预测的方法以及指标选择

目前先期进行了针对离散区间处理的讨论

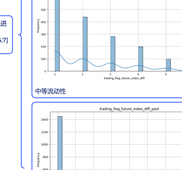
- 具体方法：我们根据与目标交易记录相似的5k（k = 80）条，将“下一笔交易发生在X个交易日后”分为(1,2](2,4](4,7](7,14](14,+∞) 五个区间，计算这k条数据落在相应区间段的概率，如果目标落在了概率最大的区间里，则标记为1，以占总行数的比例计算最终概率

结合之前的结果，对于预测我们显然需要分类进行讨论，目前基于未来7天内交易天数进行分类，(0,1]为低流动性，(2,4]为中等流动性，[5,7]为高流动性



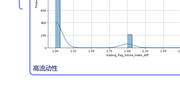
低流动性

- 对于低流动性部分，预测准确率来到了85%以上
- 但80%中几乎全部都落在了14天以上这个区间内



中等流动性

- 对于中等流动性部分，预测准确率在40%左右
- 可能的分布情况：gamma分布用于进一步拟合



高流动性

- 对于高流动性部分，预测准确率来到了80%以上
- 类似地，预测均落在了1天后就交易这个区间里

现有问题