高频算法交易与市场微观结构

笔记本：	量化研究		
创建时间：	2020/2/12 14:10	更新时间：	2020/3/3 21:22
作者：	Kiki		

1. 基础设施

- 数据建设：沪深1分钟（已有落库数据可使用），LEVEL1/LEVEL2高频数据计算价格波动因子，交易量流动性因子，成交类统计因子，订单簿因子，久期类因子，K线因子以及其他衍生因子等分析市场微观结构，成交规律，交易策略等。各类因子具体的指标参考附录。
- 平台建设：高频策略回测，目前金工平台支持一分钟策略回测，tick级别尚未支持

2. 研究问题和计划

2.1. 市场规律（交易特点等），市场微观结构统计分析

2.2 市场模型（流动性、消息类）和策略，异动预测（拐点）
- 整理level 2成交和委托数据特点和分布，找测试日期中异动案例

2.3 短期（日内）价格预测（价格趋势，波动与流动暂不开始，参考DMA）：
- SCHMEA 1: calculate features with some satistics methods in the time series, and use ml(e.g. svr) or dl(e.g. 1d cnn) models to sovle clf problem.
- features types: 3s's features, raw 1 min's features, 1min's features accumulated and calculated from 3s's features
- data source: datayes. there is no information about deas and related derivative features.
- results: cv = 5, svr, micro-f1 score(0.6-0.7) for 5 mins preditions from 1 min's featuers.
- TODO: add the config for hyper parameter, and grid search; try CNN
- SCHEMA 2: input with raw tick(3s') data with extra feature engineering, then apply time-series models(RNN/CNN with 1D-cnn inputs to handle various window length of the time series features)

2.4 Market Impact：在现有的模块框架上完善支持数据，和相应的模型更新
- 主要的Almgren 和 Kissell 框架：
- 相关因子：交易速度ν，波动率σ，买卖不均衡Q，估计的日成交量ADV，订单数量X，参与率X/ADV，流动性因子（流动股本/日成交量）Θ/ADV，交易量久期X/VT
- 由permenent_impact 和 temp_impact, 和real_impact.为线性关系
- 因子不是线性关系，比较有效的是power，而且不是sqart,因此上述因子可以口占增加log 因子；X/ADV为近似线性关系，指数接近于1；

- 增加level 2(如深交可以近似获得LOB数据)：
    - 有些因子能准备计算，如不平衡量Q，如用分钟数据估算可能会有较大误差

- 从LOB数据从微观程度进一步分析：如实际上带来market impact cost的是market oder,而非limit order；根据具体LOB数据，可以计算出基本每个order的实际market impact cost;
- 增加市场微观结构因子，如bid-ask spread, depth 和 width 等
- 对于模型训练或拟合需要的label MI可以准确计算，而不是只通过$p_0 - P^{hat}$近似
- 训练的样本，可以通过市场的成交明细产生，或者通过内部实际客户成交数据生成
- 因子测试和模型构建：先对上述的基本因子（log因子）以及扩展的微观因子进行测试分析；对有效的因子构建模型（已搭建基本tf框架），初步沿用研究框架中的因子非线性关系，而不同impact部分为线性关系进行构建模型，对模型进行训练
- 产出/产品:
  - 客户报告： 评估以及预测：评估在于交易汇总报告对MI的计算（如有L2数据，可以根据实际盘口的流动性消耗以及具体订单类型做准确计算；如果没有，例如上交，则大致用$P0 - P^{hat}$计算近似）；预测可纳入交易规划报告中，使用训练的模型进行预测；同时纳入网页报告和导出报告，已与信息中心对接，模型作为模块对接报告中，已初步沟通接口。
  - 研究报告
  - 模块上线（已和信息中心做过对接），或者技术模块发布和开源

3. 附录

- 3.1 子类因子指标
- 3.1.1 价格类因子
  - retVar:分钟收益方差

    $$\sum_i^N r_t^2$$
  - retSkr: 分钟收益偏度

    $$\frac{\sqrt{\sum_i^N r_t^3}}{RDV_t^{2/3}}$$
  - retKur:分钟收益的峰度

    $$\frac{\sqrt{\sum_i^N r_t^4}}{RDV_t^2}$$
  - retAc:分钟收益率自相关性，参数：windows_len
  - corrBc: 日内计算参数：windows_len; 日间：分钟收益率与大盘（行业）的pearson相关系数,去交易日T,股票s的日内分钟收益序列，计算收益曲线pearson相关系数，对相关系数进行纵向标准化，将烧过阈值（如3倍标准差）定义为异常数据，理解为异动资金在盘中进出所造成的影响
  - retBc: 基准（指数、行业）的分钟简单收益率
  - late_mkt_price_l1:实时价格
  - diff_mkt_vwap_l1:实时价格与均价之前跟踪误差
  - amp_price_m: 日内分钟的价格振幅
  - amp_ret_m:日内分钟的收益振幅
  - mid_price_l1
  - mkt_depth
  - ac_ret_l1: 简单收益或对数收益的某一频度的自相关性，如1min，同retAc

    $$\frac{\sum_{t=p+1}^{T} [R_t - E[R])(R_{t-p} - E[R])]}{(\sum_{t=p+1}^{T} (R_t - E[R]))^{1/2} (\sum_{t=p+1}^{T} (R_{t-p} - E[R]))^{1/2}}$$
  - price jump

- price_mom:ref DMA P263
- 3.1.2 流动性因子
  - bid_ask_spread
  - market_depth
  - market_resilience
  - price_sensitivity_λ: sensitivity to block transactions, coefficient λ from OLS estimation of the regression

    $\Delta P_t = \alpha + \lambda NVOL_T + \epsilon$, ΔP_t is the change in market price due to the market impact of orders, and $NVOL_t$ is the diff between buy and sell market depths in period t.
  - illiquidity_ratio: messure the degree of market illiquidity, D_t is the num of trades, $r_{d,t}$ is the relative price change of the following trade d during t, and $v_{d,t}$ is the trade quantity within trade d

    $\gamma_t = \frac{1}{D_t} \sum_{d=1}^{D_t} \frac{r_{d,t}}{v_{d,t}}$
  - freq_bid_quotes_l2: λ^b,e.g. could be 5 bid quotes per min
  - freq_ask_quotes_l2: λ^a,e.g. could be 5 ask quotes per min
  - chg_freq_bid_quotes_l2: $\Delta\lambda^b$,freq diff of the last min and previous min
  - chg_freq_ask_quotes_l2: $\Delta\lambda^a$,freq diff of the last min and previous min
  - ret_risk_aversion: γ
  - reservation_prices: r_a, r_b; ref: practical p138
  - market_resiliency/diversity:ref DMA ch8 and 9
  - market_cap
  - turnover
- 3.1.3 成交类因子/交易行为因子
  - amp_trade_l2:逐笔成交的振幅
  - vol_per_tran_l2:平均每单数量
  - ratio_lorder_l2: 大单比例
  - ratio_sorder_l2:小单比例
  - proportion_limit_orders_l2
  - inter_trade_dur_l2: shorter inter-trade duration,higher likelihood of unobsered gd news
  - mean/std/ske/kur_hourly_inter_trade_l2:
  - freq_buy_orders_l2:
  - buy_driven_lst_n_orders_l2:
  - averge_trade_size_l2:
  - prior_trade_aggressive:
  - s_indicator_m: s指标，以下用于构建 聪明钱因子，从分钟行情数据的加量信息，识别机构参与交易的多寡。

    $S_t = |R_t / \sqrt{(V_t)}|$
  - q_indicator_m:聪明钱因子，将分钟的S指标排序，取累计成交占20%的分位，计算两部分的vwap比值

    $Q = VWAP_{smart} / VWAP_{all}$

- kquan_m_high: 逐笔成交金额的k分位数，如每日的逐笔成交金额的16分位，在反转因子中，该指标辅助判断当日为反转还是动量属性
- 3.1.4 订单簿因子
  - bid_ask_spread_l1

    $p_i^{ask} - p_i^{bid}$
  - bid_ask_mid_price_l1

    $(p_i^{ask} + p_i^{bid})/2$
  - mean_prices_l1

    $$\frac{\sum_i^n p_i^{ask}}{n}$$
    $$\frac{\sum_i^n p_i^{bid}}{n}$$
  - mean_vols_l1

    $$\frac{\sum_i^n v_i^{ask}}{n}$$
    $$\frac{\sum_i^n v_i^{bid}}{n}$$
  - accu_diff_l1

    $$\sum_i^n p_i^{ask} - p_i^{bid}$$
    $$\sum_i^n v_i^{ask} - v_i^{bid}$$
  - dev_price_l1:price derivatives

    $$\frac{dp_i^{ask}}{dt}$$
    $$\frac{dp_i^{bid}}{dt}$$
  - dev_vol_l1:volume derivatives

    $$\frac{dv_i^{ask}}{dt}$$
    $$\frac{dv_i^{bid}}{dt}$$
  - eff_bid_ask_spread_l2:

    $$\frac{2*(mkt-mid)}{mid}$$
  - prop_mkt_orders:aggressiveness of trades
  - diff_trade_num_bid_ask: order flow
  - diff_share_num_bid_ask:order flow
  - average_spread
  - relative_spread
- 3.1.5 久期duration(价格每变动n点或交易量每增加n所需要的时间，或其他要素变动与时间关系)

  -price_duration_l2: time interval between subsequent price change of a specified magnitude
- 3.1.6 K线
  - abs_ret2vol:单一K线资金的激进程度

    $abs(R_t)/\sqrt{vol}_t$
- 3.1.7 衍生因子
  - 分钟的价格、波动进行聚类
  - average intensity of each type:ref to "Modeling high_frequency limit order book dynamics"

    $\lambda_{\Delta t}^{lb}, \lambda_{\Delta t}^{ma}, \lambda_{\Delta t}^{mb}, \lambda_{\Delta t}^{ca}, \lambda_{\Delta t}^{cb}$

    -relative intensity indicators

$$I_{\lambda_{\Delta t}^{lb} > \lambda_{\Delta T}^{lb}}$$
and etc
- accelerations(market/limit)

$$d\lambda^{ma}/dt, \, d\lambda^{lb}/dt, \, d\lambda^{mb}/dt, \, d\lambda^{la}/dt$$

- 3.1.8 集合竞价因子
  - 隔夜涨跌幅
  - 第一阶段涨幅
  - 第二阶段涨幅
  - 集合竞价成交金额占比
  - 第一阶段是否涨停
  - 第二阶段是否涨停
  - 第二阶段价格是否平稳上升
  - 第一阶段委比变化（（9：20委比-9：15委比）/9:15委比）
  - 第二阶段委比变化（（9：25委比-9：20委比）/9：20委比）
  - 集合竞价阶段成交金额占当天成交金额比例？？未来函数
  - 第二阶段的委买一价，委卖一价均值的平均值
  - 第二阶段的委买一价，委卖一价均值的最大值
  - 第二阶段的委买一价，委卖一价均值的最小值
  - 第二阶段的委买一价，委卖一价均值的变化比率
- 3.2 市场规律、行为数据验证
  - proportion of limit orders used by a particular trader or trading strategy -> confidence in their information->used to extracting additional info from trading decisions and order flow of the trader
  - 筛选大单作为信号
  - bid-ask spread 和流动性和transaction cost 关系
  - bid-ask spread 预测模型，practical p120
  - tick data 到达模型，practical p121, 泊松分布
  - optimal limit bid/ask price, ref:practical P139

    $$b = r^b - \frac{1}{\gamma} ln(1 - \lambda \frac{\lambda^b}{\Delta \lambda^b})$$
    $$a = r^a - \frac{1}{\gamma} ln(1 - \lambda \frac{\lambda^a}{\Delta \lambda^a})$$

  - market breadth and depth of the limit order book predicts 30 percents of the impending price moves, depth of the limit order book can forecast future volatility of asset prices; limit orders carry private info around events; ref practical p 142
  - information-based impact;adverse seletion components of the bid-ask spread;probability of informed trading; practical p147
  - price changes: difference between the bid-ask mid-points at times t and t+n,ref practical p159

    $$Diff_t = \beta_0 + \beta_1 Size_t + \beta_2 Aggressiveness_t + \beta_3 Institutional_t + \gamma_1 D_{1\,t} + ... + \gamma_k D_{k-1,\,t} + \epsilon_t$$

    t is the time of the order submission, n=5,then 60 mins after order submission. Size is the num of shares in the particular order divided by the mean daily volume of shares traded in the particular stock over the sample period. for buy orders,aggressiveness is a dummy that takese the value 1 if the order is placed at or better than the standing quote and zero otherwise. Institutional is a dummy variable that takese the value 1 for institutional

orders and 0 for individual orders. D_1 to D_{k-1} are stock specific dummies associated with the particular stock that was traded.

- market aggressiveness show autocorrelation by:
  - large institutional orders that are transmitted in smaller sliced over an extened period of time at comparable levels of market aggressiveness
  - simple price momentum
- order flow prediction model, ref practical p163
- 逐笔交易金额的分布，日内逐笔交易的平均数大概在80分位数，通过日内逐笔成交金额为筛选，将交易日分类high和Low两个类别，high的组别里面呈现反转，而Low则是动量
- a common application is identifying turning points in the time series, to determin when a sudden change in direction is likely
- frequency of buy orders and the 5-mins' return
- positive serial correlation for the arrival of order types(e.g. buy followed by another buy)
- APM因子：因子目标：度量上午与下午的差异

  记录每日上午和下午的股票收益率 $r_i^{am/pm}$,指数收益率为 $R_i^{am/pm}$

  -对股票和指数收益做回归

  $r_i = \alpha + \beta * R_i + \epsilon_i$

  - 计算上午和下午的残差差值

  $\delta_t = \epsilon_t^{am} - \epsilon_t^{pm}$

  - 度量上午与下午残差的差异程度

  $stat = \frac{\mu(\delta_t)}{\sigma(\delta_t)/N^{1/2}}$

  - 对动量因子进行截面回归 后的残差 ϵ_j

  $stat_j = bRet20_j + \epsilon_j$

- 测试集模块
  1. 每个行业龙头，主要市值排行
  2. 每种规模，如300，500，800，1000，中小板，创业板
  3. 流动性看：成交量活跃度，按换手率，和成交量日间日内波动
  4. 异动股，日间（短期）、日内
  5. 风格：动量、反转
  6. 对每种指数，300，500