Haute Frequency Trading

Inhaltsverzeichnis

Introduction

Microstructure effect: At the microscopic scale, prices are very different from Brownian type sample paths. Therefore, intricate to model.

- ⇒ Model is a **good model** if the assumptions on the market participants are reasonable, leading to acceptable market features and mostly if it enables to draw relevant conclusions or recommendations. If the case is not motivated by agents' behaviour, a model is good if it reasonably reproduces the stylized facts of the main quantities of interest and, mostly, if it is useful for market practitioners.

Scales of resolution for models

L0- The ultimate level of the order book: In continuous time with discrete values in an appropriate state space which describes all the events of a limit order book.
L1- The ultra-high frequency level for the price: Model all transaction prices and durations between these transactions.
L2- Intermediate high frequency level: Here one does not focus on durations but essentially on the price, regularly sampled, for example every second or every minute.
L3- The macroscopic level: Price is viewed as a continuous semi-martingale. It is the dominant and historical approach.

- ⇒ Best, but also most difficult, approach would be to satisfy all levels. Normally just L2 to L3. Besides, question of which is the modeled price- last traded, mid-quote, best-bid…

Signature plot

Idea: We want to estimate the volatility. Normally it works well to use the sum of squared returns (these are converging to the cumulative volatility). For the discrete case, we get the signature plot as following.

We assume to have price observations P_t at times $t = \frac{i}{n}, n \in N, i = 1, \dots, n$. The signature plot associates to $k = 1, \dots, n$:

$$RV_n(k) = \sum_{i=0}^{\lfloor n/k \rfloor - 1} (P_{k(i+1)/n} - P_{ki/n})^2$$

If P_t is a continuous semi-martingale, as soon as k/n is large enough, the signature plot is close to the quadratic variation. However, very often a decreasing functional, stabilizing for sampling periods larger than 10 minutes (depending on the asset of course).
But, not directly applicable for high frequency.

From low to high frequency

Starting from a continuous semi-martingale (efficient/latent price) and apply stochastic mechanism to derive the observed prices at higher frequencies.

- ⇒ Mechanism must be negligible in low frequencies to stay close to the macroscopic models.
- ⇒ Efficient price exists in a sense of market agreement, but one does not have to be convinced that it makes sense at these very fine scales. Even if it exists only as a construct to build the model.

Going from L3 to L2

Additive microstructure noise approach: We will add some noise to the macrostructure model. P_t is the model price at time t, X_t the semi-martingale and ϵ_t the microstructure noise term:

$$\epsilon_t = \log P_t - \log X_t$$

More precisely, for discrete observations

$$\log P_{i/n} = \log X_{i/n} + \epsilon_i^n$$

It is additive, if the noise term is centered, stationary and independent of X.
Problem: Models an unobservable quantity (microstructure noise term).
Properties wanted for L1: The model does not satisfy any of the following properties (which are wanted for L1 and therefore it cannot be extended):

(i) Observed prices are discrete, which is true since market prices usually stay on a tick grid, but not for continuous semi-martingales (therefore noise necessary).
(ii) For many assets, quick oscillations of transaction prices between two values (bid-ask bounce).
(iii) $(logP)_t$ is almost surely finite (since P_t is a jump process with finite number of jumps during a bounded interval

But are good for modeling scales about 5 minutes.
Rounded models
Efficient price modeled by a continuous martingale X_t and observed price given by

$$(X_{i/n}^{(\alpha_n)}, i = 0, \dots, n) \text{ where } X_{i/n}^{(\alpha_n)} = \alpha_n \lfloor X_{i/n}/\alpha_n \rfloor$$

Here, α_n is the rounding error corresponding to the tick size.

- ⇒ Prices are discrete and behave as observations from a continuous semi-martingale in the low frequencies, the rounding effect becoming negligible. Furthermore, Properties (ii) and (iii) can be satisfied.

Problem: Can still not be extended to L1. If for any time t the observed price is given by the rounding value of a semi-martingale. This leads to an observed price with an infinite number of jumps on a finite interval, which is of course hardly acceptable.
Suitable properties for L3 -> L1
Properties of the model

- Model for prices and durations (in particular, no notion of sampling frequency is required)
- Testable, interpretable, and useful (e.g., for building statistical procedures)
- Usual stylized facts of returns, durations, and volatility (in a loose sense here). Inverse relation between durations and volatility.

Prices

- Continuous semi-martingale type behavior at large sampling scales.
- A clear definition of the price.
- Discrete prices.
- Finite quadratic variation for the price.
- Bid-Ask bounce.

Going from L1 to L3

Fine coarse models: Seem more natural (going from a small scale going to a large one).
Problem: Built without any reference to some kind of efficient price. Mathematical, no "more natural" since also L3 to L1 models use suitable stochastic processes.

Optimal Portfolio Liquidation

We want to sell a large quantity of a stock (or of several stocks) in one day. How to choose a good way to split this large order in time and volume?
Naïve strategies: Sell everything immediately (huge, but known, transaction costs due to the influence on the order book). Sell regularly in small amounts (little transaction costs, but volatility risk and thus, unknown outcome).

Almgren and Chriss Model

Set up: We have X shares of one asset at time $t_0 = 0$ and want to sell till T. We split our interval [0, T] into N intervals of length $\tau = T/N$ and $t_k = k\tau$. A trading strategy is a vector $(x_0, \dots, x_N)$ with x_k number of assets we have at time t_k. We set $n_k = x_{k-1} - x_k$ number of assets sold between t_{k-1} and t_k.
Price movements:

- Drift and volatility
- Market Impact
  - Permanent: Participants see us selling large amounts and correct their prizes down. The equilibrium price is thus corrected permanently:
    $$S_k = S_{k-1} + \sigma\tau^{\frac{1}{2}}\xi_k - \tau g(n_k/\tau)$$
    with ξ_k iid. standard Gaussian and n_k/τ average trading rate between t_{k-1} and t_k
  - Temporary: By selling assets, we are liquidity taker. If we sell a large amount of shares, our price per share is significantly worse than when selling only one share. We assume this effect is temporary and the liquidity comes back after each period. We set $\widetilde{S_k} = \frac{\sum n_{k,i} p_i}{n_k}$ with $n_{k,i}$ number of shares sold at price p_i between t_{k-1} and t_k. We set:
    $$\tilde{S}_k = S_{k-1} - h(n_k/\tau)$$
    Here the term $h(n_k/\tau)$ does not influence the next equilibrium price S_k

Profit and Loss: The result of the sale of the asset:
$$\sum_{k=1}^{N} n_k \tilde{S}_k = XS_0 + \sum_{k=1}^{N} (\sigma\tau^{\frac{1}{2}}\xi_k - \tau g(n_k/\tau))x_k - \sum_{k=1}^{N} n_k h(n_k/\tau)$$

Then the trading costs C are $XS_0 - \sum_{k=1}^{N} n_k \widetilde{S_k}$ and these are equal to the sum of volume costs, permanent impact costs and temporary impact costs. We want to minimize for a risk aversion parameter λ:

Assumptions

We assume a linear permanent impact $g(v) = \gamma v$. Thus, we have for the permanent impact:

$$S_k = S_0 + \sigma \sum_{j=1}^{k} \tau^{\frac{1}{2}} \xi_j - \gamma(X - x_k)$$

We assume an affine temporary impact $h(v) = \epsilon + \mu v$, where ϵ represents a fixed cost (fees + bid-ask spread).

Naïve strategies

Regular strategy: Take $n_k = \frac{X}{N}$. We can show that this strategy has the smallest expectation for the transaction costs, but high variance if T is large.
Selling everything immediately: Take $n_1 = X$ and $n_2 = \cdots = n_N = 0$. This strategy has the lowest variance. However, if τ (length of interval) is small, the expectation can be high.

Optimization

We want to optimize:

$$\mathbb{E}(C) + \lambda \mathbb{V}(C) = \frac{1}{2}\gamma X^2 + \epsilon X + \frac{\mu - \frac{1}{2}\gamma\tau}{\tau} \sum_{k=1}^{N} (x_{k-1} - x_k)^2 + \lambda\sigma^2 \sum_{k=1}^{N} \tau x_k^2$$

Hence, we build the derivates for $x_1, \ldots, x_{N_1}$ and get with K satisfying $\frac{2}{\tau^2}(cosh(K\tau) - 1) = \frac{\lambda\sigma^2}{\mu - \frac{1}{2}\gamma\tau}$ the following solution:

$$x_j = \frac{\sinh(K(T - t_j))}{\sinh(KT)} X$$

$$n_j = \frac{2\sinh(K\tau/2)}{\sinh(KT)} \cosh(K(T - j\tau + \tau/2)) X$$

Final Remarks:

- Solution is time-homogenous (if we compute the optimal strategy in t_k, we obtain the values between t_k and T of the optimal strategy computed in t_0.
- Optimal trajectories are very sensitive to the volatility. It is therefore important to obtain accurate volatility estimates.
- Can be extended to n dimensions (several assets), then we also need correlation values.

But we get only a time window, not the exact moment.

Model with uncertainty zones

In an idealistic framework, transactions would occur when the efficient price crosses the tick grid. In practice, uncertainty about the efficient price and aversion for price changes of market participants. The price changes only when market participants are convinced that the efficient price is far from the last traded price. We introduce a parameter μ quantifying this aversion for price changes.

Interpretation of μ

In ultra-high frequency, the order book cannot follow the efficient price and is reluctant to price changes. Reluctancy measured by μ.

⇒ 2 μα represents the implicit spread of a large tick asset.

⇒ A small $\mu < \frac{1}{2}$ means that for market participants, the tick size is too big and conversely.

Notation

X_t : Efficient price
α : Tick size
t_i : Time of the i-th transaction with price change
P_{t_i} : Transaction price at time t_i
Size of the i-th price jump $L_i = \frac{|P_{t_{i+1}} - Pt_i|}{\alpha}$
χ_t : Explanatory variables process for the size of the jumps
Uncertainty zones $U_k = [0,\infty) \times (d_k, u_k)$ with
$d_k = \left(k + \frac{1}{2} - \mu\right)\alpha$ and $u_k = \left(k + \frac{1}{2} + \mu\right)\alpha$
τ_i : i-th exit time of an uncertainty zone

Dynamics

$d\log X_u = a_u du + \sigma_{u-} dW_u$
$\mathbb{P}[L_i = s|\mathcal{F}_{\tau_i}] = \phi_s(\chi_{\tau_i})$
$\tau_{i+1} = \inf\{t > \tau_i, X_t = X_{\tau_i}^{(\alpha)} \pm \alpha(L_i - \frac{1}{2} + \mu)\}$
with $X_{\tau_i}^{(\alpha)}$ the value of X_{τ_i} rounded to the nearest multiple of α
$\tau_i \leq t_i < \tau_{i+1}$ and $P_{t_i} = X_{\tau_i}^{(\alpha)}$

Volatility Estimation

Efficient price

$$X_{\tau_i} = P_{t_i} - \alpha\left(\frac{1}{2} - \mu\right)\text{sign}(P_{t_i} - P_{t_{i-1}})$$
$$\hat{X}_{\tau_i} = P_{t_i} - \alpha\left(\frac{1}{2} - \hat{\mu}\right)\text{sign}(P_{t_i} - P_{t_{i-1}})$$

We use here an estimation of μ. μ describes the distribution of high frequency tick returns:

⇒ Small: Tick size large. (Uncertainty zone small, which means that we have a strong (average) reversion of the observed price and thus, a decreasing signature plot.)

⇒ $\mu \sim 1/2$: Tick size in some sense optimal. (Last traded price can be seen as a Brownian Motion, thus no microstructure effects. Flat signature plot and uncertainty zone = 1 tick.)

Estimation of μ

Idea

A continuation/alternation is a price variation whose direction is the same/opposite as the one of the preceding variations. We set N^c the number of continuations and N^a the number of alternations. We get an estimator for μ:

$$\hat{\mu} = \frac{N^c}{2N^a}$$

Details

Let $N_{\alpha,t} = \#\{t_i, t_i < t\}$ number of transactions with a price change till t. We regard now, the number of a specific price change between two exits of uncertainty zones:

$$N^{(\cdot)}_{\alpha,t,k} = \sum_{i=1}^{N_{\alpha,t}} \mathbb{I}_{\{|X_{\tau_i} - X_{\tau_{i-1}}| = \cdot\}}$$

We will consider $c = \alpha k$ and $a = \alpha(k - 1 + 2\mu)$. Finally, we have as an estimation for μ

$$\hat{\mu}_t = \sum_{k=1}^{m} \lambda_{\alpha,t,k} u_{\alpha,t,k}$$

where λ describes the proportion of our considered price changes (summed) to all possible price changes, and u portrays the proportion between the considered price changes rescaled:

$$\lambda_{\alpha,t,k} = \frac{N^{(a)}_{\alpha,t,k} + N^{(c)}_{\alpha,t,k}}{\sum_{j=1}^{m} N^{(a)}_{\alpha,t,k} + N^{(c)}_{\alpha,t,k}} \text{ and } u_{\alpha,t,k} = \frac{1}{2}\left[k\left(\frac{N^{(c)}_{\alpha,t,k}}{N^{(a)}_{\alpha,t,k}} - 1\right) + 1\right]$$

Estimation of the volatility

We consider:

$$\hat{RV}_t = \sum_{i=1}^{N_{\alpha,t}} (\log(\hat{X}_{\tau_i}) - \log(\hat{X}_{\tau_{i-1}}))^2$$

Then we have as a convergence

$$\alpha^{-1}(\hat{RV}_t - RV_t) \to \gamma_t \int_0^t v_u dW_{\theta_u}$$

where W is a Brownian motion independent of B and θ_u; γ_u and v_u depend on X_u, σ_u and explanatory variables, involving for example the order book.

Estimation of covariance

We consider two assets with the following dynamics:

$$d\log X_t = \mu_t^X dt + \sigma_{t-}^X dW_t$$

$$d\log Y_t = \mu_t^Y dt + \sigma_{t-}^Y dB_t$$

With a correlation between the two Brownian motions

$$d\langle W, B\rangle_t = \rho_t dt$$

We want to estimate the integral

$$\int_0^1 \rho_t \sigma_t^X \sigma_t^Y dt$$

We have the regular estimator with accuracy $\frac{1}{\sqrt{n}}$

$$\hat{c_n} = \sum_{i=1}^{n} \Delta_i^n X \Delta_i^n Y \text{ with } \Delta_i^n X = \log X_{i/n} - \log X_{(i-1)/n}$$

Problems: Asynchrony of data and microstructure effects are not considered in the regular estimator.

If volatility and correlation are constant, we get as an estimator with accuracy $\frac{1}{\sqrt{n}}$:

$$\frac{\hat{c}_n}{\sqrt{\sum_{i=1}^{n}(\Delta_i^n X)^2 \sum_{i=1}^{n}(\Delta_i^n Y)^2}}$$

$$\hat{\rho}_n = \frac{2}{\pi}\frac{\hat{c}_n}{\hat{a}_n} \text{ with } \hat{a}_n = \sum_{i=1}^{n-1} |\Delta_{i+1}^n X \Delta_i^n Y|$$

If they are not constant, we get as an estimator:

Previous tick scheme

If we assume, that we have just some observed prices at times $T^{\cdot,i}$

$$\bar{X}_t = X_{T^{X,i}} \text{ for } t \in [T^{X,i}, T^{X,i+1})$$

$$\bar{Y}_t = X_{T^{Y,i}} \text{ for } t \in [T^{Y,i}, T^{Y,i+1})$$

Then we get the modified estimator for a given h
Hayashi-Yoshida estimator
Advantage: Does not need any selection of h and is convergent if the arrival times are

$$\sum_{i=1}^{m}(\log \bar{X}_{ih} - \log \bar{X}_{(i-1)h})(\log \bar{Y}_{ih} - \log \bar{Y}_{(i-1)h})$$

$$I_i^{\cdot} = (T^{\cdot,i}, T^{\cdot,i+1}]$$

$$U_n = \sum_{i,j} \Delta X(I_i^X)\Delta Y(I_i^Y)\mathbb{I}_{\{I_i^X \cap I_j^Y \neq \emptyset\}}$$

independent from the price.
Problem: Not robust to microstructure effects.

$\Rightarrow$ Consistent estimator for the covariation (by using estimated values of the efficient prices)

Lead-Lag Estimation

Some assets are leading some other assets. This means that a “lagger" asset may partially reproduce the behavior of a “leader" asset (with some time delay). We want to estimate that time shift.

Bachelier Model

For $t \in [0,1]$ and $(B^{(1)}, B^{(2)})$ such that $\langle B^{(1)}, B^{(2)} \rangle_t = \rho t$

$$X_t := x_0 + \sigma_1 B_t^{(1)},\ \tilde{Y}_t := y_0 + \sigma_2 B_t^{(2)}$$

We define $Y_t = \widetilde{Y_{t-\theta}}$ for $t \in [\theta, 1]$. Our lead-lag model is given by the process (X_t, Y_t) and we have:

$$X_t = x_0 + \sigma_1 B_t^{(1)}$$

$$Y_t = y_0 + \rho\sigma_2 B_{t-\theta}^{(1)} + \sigma_2(1-\rho^2)^{1/2} W_{t-\theta}$$

Estimation

We assume, that the data arrives at regular and synchronous time stamps:

$$(X_0, Y_0), (X_{\Delta_n}, Y_{\Delta_n}), (X_{2\Delta_n}, Y_{2\Delta_n}), \ldots, (X_1, Y_1)$$

And suppose $\theta = k_0\Delta_n, k_0 \in Z$
We define:

$$C_n(k) := \sum_i (X_{i\Delta_n} - X_{(i-1)\Delta_n})(Y_{(i+k)\Delta_n} - Y_{(i+k-1)\Delta_n})$$

Then we can maximize $C_n(k)$ in k, by that we detect the value k_0 that defines in that case.

Lead-Lag Model

Let $\theta > 0$ and set $F^{\theta} = (\mathcal{F}_{t-\theta})_{t\geq 0}$. We consider:

$$X = X^C + A,\ Y = Y^C + B$$

Where $(X_t^C)_{t\geq 0}$ is a continuous $\mathcal{F}$-local martingale and $(Y_t^C)_{t\geq 0}$ is a continuous F^{θ}-local

$$\exists v_n \to 0 : \frac{\max\{\sup|I_i^X|, \sup|I_i^Y|\}}{v_n} \to 0$$

martingale. We suppose that:
Besides $T^{X,i}$ are F^{v_n}-stopping times, and $T^{Y,i}$ are $F^{\theta+v_n}$-stopping times.

Estimation

We set for $(I_j^Y)_{-\theta} = (T^{Y,j} - \theta, T^{Y,j+1} - \theta]$:

$$U_n(\theta) = \sum_{i,j} \Delta X(I_i^X)\Delta Y(I_j^Y)\mathbb{I}_{\{I_i^X \cap (I_j^Y)_{-\theta} \neq \emptyset\}}$$

For a sufficiently fine grid $\mathcal{G}^n$, the estimator $\widehat{\theta_n}$ which is the solution for

$$|U_n(\hat{\theta}_n)| = \max_{\theta \in \mathcal{G}^n} |U_n(\theta)|$$

Tick Values and Regulation

Exchange rule: There exists a price grid for orders.
Tick size: Smallest Price Increment.
Tick value: how much money is lost or gained per contract per tick move.
Large tick asset: such that the bid-ask spread is almost always equal to one tick (small tick assets have spreads that are typically a few ticks).

- ⇒ In literature: relationships between spread and market quantities for small tick assets, but not for large ones.

Spread Theory for Small Tick Assets

Madhavan, Richardson, Roomans economic model (MRR model)

p_{i+1}: ex post true/efficient price after the i-th trade (all transactions have the same volume).
ϵ_i: sign of the i-th trade.
ξ_i: independent, centered shock component with variance v^2.
Then the MRR model is given by

$$p_{i+1} - p_i = \xi_i + \theta\epsilon_i$$

Market makers post (pre trade) bid and ask prices ai and bi given by

$$a_i = p_i + \theta + \phi, b_i = p_i - \theta - \phi$$

with ϕ an extra compensation claimed by market makers, covering processing costs and the shock component risk. If it is zero, the average cost of a market order is zero (= $p_{i+1} - a_i, p_{i+1} - b_i$).

Bid-Ask Spread

$$S = a - b = 2(\theta + \phi)$$

The news component is negligible (literature). Thus, we get for the variance per trade of the efficient price:

$$\sigma_1^2 = E\left((p_{i+1} - p_i)^2\right) = v^2 + \theta^2 \sim \theta^2$$

And we have finally

$$S = 2(\sigma_1 + \phi)$$

Market Making Strategy

Market makers: patient traders. Send limit orders (thus, delayed execution). Get the spread but has volatility risk.
Market takers: impatient traders. Send market orders (thus, immediate execution). Pay the spread but no volatility risk.

Wyart et al. Approach

Average cost of a market order

$$P\&L = \frac{S}{2} - \frac{c}{2}\sigma_1$$

Idea: Any agent can choose between market orders and limit orders.

- ⇒ Both types of orders will have the same average (ex post) cost = 0.
- ⇒ Market makers' P&L = 0 (if not so, another market maker comes with a slightly tighter spread).

And we get finally $S \sim c\sigma_1$, which is very well satisfied in market data.
Simplification: In uncertainty zones (see volatility estimation), we use $S = \alpha$ constant (1 tick).

Market Order Areas

Ask Zone (only buy market orders), Bid Zone (only sell market orders), and Uncertainty (Ask/Bid) Zone.

⇒ Distance between Bid and Ask Zone is equal to 2 $\alpha\mu$, which represents implicit, unobservable spread.

Implicit Spread and Volatility per Trade

We regard the total number of trades M. If we consider the implicit spread 2 $\alpha\mu$, we now want to study the relationship for large assets:

$$\alpha\mu = \frac{S}{2} \sim \frac{\sigma}{\sqrt{M}} + \phi$$

Market Order Costs

Average ex post cost of a market order:

$$\frac{\alpha}{2} - \mu\alpha$$

Average P&L per trade of the market makers = average cost of a market order:

$$\mu\alpha = c\frac{\sigma}{\sqrt{M}} + \phi$$

$\boldsymbol{\mu < \frac{1}{2}}$**:** Limit orders are profitable whereas market orders are costly. That is a natural state, otherwise market makers would increase the spread (decreasing signature plot).

Consequences of Tick Value Changes

α too small: encourages free-riding (directional HFT) and traditional market makers cannot fix their quotes.
α too large: implies price sloppiness. Moreover, it favors speed (race to the top of book), thus high investments in infrastructure.

Analyzing Order Book (not in 2023/24)

Introduction

Aim: Understanding the behaviors of market participants at different limits of the order book. Providing a realistic market simulator, enabling to compute execution costs of complex trading strategies.
Approach: Model the dynamics of a model book where the reference price stays constant. Then introduce dynamics of reference price.

- No individual agent.
- Intelligence is added through a mean field game type approach.
- Empirical studies of trader's average behaviors become possible.

Constant Reference Price

Assumptions:

- Reference price p_{ref} stays constant.
- K limits are considered on each side.
- At the bid side of the order book, market participants can only send buy limit orders, cancel existing buy orders, or send sell market orders.
- At the ask side of the order book, market participants can only send sell limit orders, cancel existing sell orders, or send buy market orders.

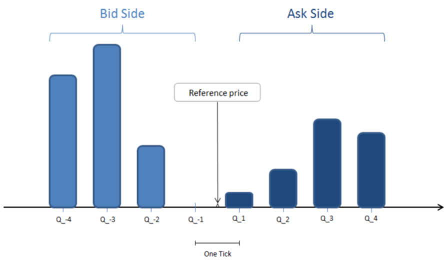
- A constant order size is assumed for each limit Q_i (different values at different limits are allowed).

Limit Order Books (LOB)

As a continuous time Markov jump processes.

- $2K$-dimensional LOB state $X(t) = (Q_{-K}(t), \dots, Q_{-1}(t), Q_1(t), \dots, Q_K(t))$, where K denotes the number of available limits on each side and $Q_.$ the limit.
- Order flow intensities:
  - Market orders $\lambda^M_{\text{buy/sell}}(x)$,
  - Limit Orders at Q_i $\lambda^L_i(x)$
  - Cancellations at Q_i $\lambda^C_i(x)$
- The associated infinitesimal generator matrix Q

$$Q_{x,x+e_i} = \lambda^L_i(x)$$
$$Q_{x,x-e_i} = \lambda^C_i(x) + \lambda^M_{\text{buy}}(x)\mathbb{I}_{\text{bestask}(x)=i} \text{ if } i > 0$$
$$Q_{x,x-e_i} = \lambda^C_i(x) + \lambda^M_{\text{sell}}(x)\mathbb{I}_{\text{bestbid}(x)=i} \text{ if } i < 0$$
$$Q_{x,x} = -\sum_{y \in \Omega, y \neq x} Q_{x,y}$$
$$Q_{x,y} = 0 \text{ otherwise}$$



$\Rightarrow$ Under reasonable assumptions: X is ergodic (invariant measure π, i.e. $\pi P = \pi$, and $\lim_{t\to\infty} P_{xy}(t) = \pi_y$ with P_{xy} transition probability from x to y).

Estimation of Reference Price

When the spread is odd (in tick units):

$$p_{\text{ref}} = p_{\text{mid}} = \frac{p_{\text{bestbid}} + p_{\text{bestask}}}{2}$$

When the spread is even, we use either:

$$p_{\text{mid}} + \frac{\text{tick size}}{2} \text{ or } p_{\text{mid}} - \frac{\text{tick size}}{2}$$

Choosing which one is closer to the previous value p_{ref}.

Independent Queue Model

We assume that the three types of orders (market orders, limit orders and cancellations) are independent flows. The Order arrivals are modeled with three birth-death processes with arrival/departure rate:

$$\rho_i(n) = \frac{\lambda^L_i(n)}{\lambda^C_i(n+1) + \lambda^M_i(n+1)}$$

Queue size tends to increase for $\rho > 1$ and to decrease for $\rho < 1$. The stationary distribution is given by:

$$\pi_i(n) = \pi_i(0)\prod_{j=1}^{n} \rho_i(j-1)$$
$$\pi_i(0) = \left(1 + \sum_{n=1}^{\infty}\prod_{j=1}^{n} \rho_i(j-1)\right)^{-1}$$

Two Sets of Dependent Queue Model

Market orders are only sent to first and second limits and consume quantities at best limits. Institutional traders and brokers tend to place most of their limit orders at best limits, while many market makers, arbitragers and other high frequency traders stand also in queues beyond these best limits. Thus, the dynamics at second limit depend also on whether the first limit is empty.

Modelling bid-ask dependences

Market participants adjust their trading rate not only according to the target queue size, but also to whether the opposite queue size is small, usual, or large.

Dynamic reference price and time consistent model (not in 23/24)

Purely Order Book driven Model

We consider price changes, triggered with probability θ by one of the following events:

- The insertion of a buy (sell) limit order within the bid-ask spread, while Q_1 (Q_{-1}) is empty at this moment.
- A cancellation of the last limit order at the best offer queue.
- A market order that consumes all the available quantity at the best offer queue.
- ⇒ Q_i becomes $Q_{i\pm1}$ dependent of the price change.

Price fluctuations totally generated by the dynamics of the order book.
The price volatility is naturally an increasing function of θ. The maximum achievable volatility (mechanical volatility) is often smaller than the empirical volatility.

Queue-Reactive Model

We also consider price changes with probability θ. Additionally, with probability θ^{reinit} the whole LOB is redrawn from its invariant distribution around the new p_{ref} when p_{ref} changes. That portrays the case, where (due to exogenous information) market participants adjust very quickly their order flows around the new p_{ref}.

- ⇒ After suitable rescaling, we obtain a diffusive behavior at large time scales for the price in the queue reactive model.
- ⇒ Can be used as a market simulator for analyzing order placement tactics.

Order Placement Tactics

Fire and forget: At t = 0, post a limit order at the best offer queue. When the mid-price changes, cancel the limit order and send a market order at the opposite side with all the remaining quantities if any. At t = T, send all the remaining quantities at the opposite side to finish the execution.

Pegging to the best: At t = 0, post a limit order at the best offer queue, and then “peg" to it: if the best offer price changes, cancel the existing order and repost all the remaining quantities at the new best offer queue. If our order is the only remaining order in the best offer queue, cancel it and repost the remaining quantities at the newly revealed best offer queue. At t = T, send all the remaining quantities at the opposite side to finish the execution.

- ⇒ Will be used with some volume scheduling tactics.
- ⇒ **Pegging to the best:** progressive process, depending both on the target quantity and the duration (better for VWAP execution benchmark).
  Fire and forget: instantaneous and depends essentially on the target quantity (slightly better for S_0 execution benchmark).

Optimal make-take fees

Problem: Exchanges are in competition. Hence, they must attract liquidity.

- ⇒ Using make-take fees system, that is charging in an asymmetric way liquidity provision and liquidity consumption.

But slower traders have no longer access to the order book (or just for unfavorable situations) and leave in time of stress.

Arrival of Market Orders

We model the arrival of buy and sell market orders by a point process $(N_t^a)_{t\geq 0}$ and $(N_t^b)_{t\geq 0}$ with intensity $(\lambda_t^a)_{t\geq 0}$ and $(\lambda_t^b)_{t\geq 0}$. The inventory of the market maker $Q_t = N_t^b - N_t^a$. Furthermore, we consider a threshold inventory $\bar{q}$ above which the market maker stops quoting on the ask or bid side. From financial arguments:

$$\lambda_t^a = \lambda(\delta_t^a) 1_{\{Q_t > \bar{q}\}}$$
$$\lambda_t^b = \lambda(\delta_t^b) 1_{\{Q_t < \bar{q}\}}$$

Where $\lambda(x) = Ae^{-k(x+c)/\sigma}$

Market Maker

He knows the efficient/mid-price of the asset with some price volatility:

$$S_t = S_0 + \sigma W_t$$

He fixes ask- and bid-prices:

$$P_t^a = S_t + \delta_t^a$$
$$P_t^b = S_t - \delta_t^b$$

The cash flow of the market maker is:

$$X_t^\delta = \int_0^t P_u^a dN_u^a - \int_0^t P_u^b dN_u^b$$

The inventory risk of the market maker is $Q_t S_t$. For a given contract ξ (random variable) the market chooses his spread δ by maximizing his utility.

Market Maker Problem

Under the exchange incentive policy ξ, the market maker solves now (under the condition $V_{MM}(\xi) > R$):

$$V_{MM}(\xi) = \sup_\delta \mathbb{E}^\delta\left[-\exp\left(-\gamma(X_T^\delta + Q_T S_T + \xi) \right) \right]$$

Exchange Problem

We assume that the exchange:

- Earns $c > 0$ for each market order occurring in its platform.
- Pays the incentive policy ξ (depend on S, N^a and N^b) to the market maker.

The profit and loss of the exchange is:

$$c(N_T^a + N_T^b) - \xi$$

The exchange designs the contract ξ by solving (η risk aversion, δ^* optimal spread):

$$V_E = \sup_\xi \mathbb{E}^{\delta^*}\left[-\exp\left(-\eta(c(N_T^a + N_T^b) - \xi) \right) \right]$$

We can represent each contract ξ under the form $\xi = Y_T$:

$$dY_t = Z_t^a dN_t^a + Z_t^b dN_t^b + Z_t^S dS_t - H(Z_t, Q_t) dt$$

Theorem

The contract $\xi^\star$ that solves the exchange problem is given by

$$\xi^\star = Y^\star + \int_0^T Z_t^{a,\star} dN_t^a + Z_t^{b,\star} dN_t^b + Z_t^{S,\star} dS_t - H(Z_t^\star, Q_t) dt,$$

with

$$Z_t^{a\star} = -\frac{\sigma}{k} \log\left(\frac{u(t, Q_t)}{u(t, Q_t - 1)} \right) + \hat{c},$$

$$Z_t^{b\star} = -\frac{\sigma}{k} \log\left(\frac{u(t, Q_t)}{u(t, Q_t + 1)} \right) + \hat{c}, \quad Z_t^{S\star} = -\frac{\gamma}{\eta + \gamma} Q_t.$$

with $\hat{c} = c + \frac{1}{\eta} \log\left(1 - \frac{\sigma^2 \gamma \eta}{(k + \sigma\gamma)(k + \sigma\eta)} \right)$.

Some Remarks

- When the inventory is highly positive, the exchange provides incentives to the market-maker so that it attracts buy market orders (and tries to dissuade him to accept more sell market orders).
- The exchange transfers the totality of the taker fee c to the market maker. It is neutral to the value of c. But c plays important role in the optimal spread (fix transaction cost c so that the spread is close to one tick)

Conclusions

Benefits of exchange incentive policy

- Smaller spreads.
- Better market liquidity
- Increase of the profit and loss of the market maker and exchange
- Lower transaction costs.

Volatility is Rough

Introduction

Three classes of volatility

1) Deterministic Volatility
2) Local Volatility
3) Stochastic Volatility

Classical volatility models generate reasonable dynamics, but not allow to fit the volatility surface, particularly the ATM-skew. The skew is well approximated by a power-law function of time to expiry τ. But conventional models generate a term structure of ATM skew that is constant for small τ.

Therefore, we introduce the fractional Brownian Motion.

The **Fractional Brownian Motion** with Hurst parameter H is the only process who satisfies:

$(W_{at}^H) \sim a^H (W_t^H)$ (Self-similarity)

$(W_{t+h}^H - W_t^H) \sim W_h^H$ (Stationary increments)

Gaussian process with $\mathbb{E}[W_1^H] = 0$ and $\mathbb{E}[(W_1^H)^2] = 1$

The Hurst parameter determines, *how* rough the volatility is. Models where the volatility is driven by a fractional Brownian Motion generate an ATM volatility skew of the form $\tau^{H-1/2}$, at least for small τ.

Rough FSV Model

The starting point is to consider the scaling of the moments of the increments of the log-volatility:

$$m(\Delta, q) = \mathbb{E}[|\log(\sigma_{t+\Delta}) - \log(\sigma_t)|^q]$$

The behavior of $m(\Delta, q)$, when Δ is close to zero, is related to the smoothness of the volatility. Empirical findings suggest, we model the log-volatility as a fractional Brownian Motion:

$$\sigma_t = \sigma e^{\nu W_t^H}$$

An important property of volatility time series is their multi-scaling behavior, i.e. one observes essentially the same law whatever the time scale. There are periods of high and low market activity at different time scales (very few models reproduce this property, see multifractal models).

Kommentiert [NR1]: increments of log-volatility of various assets enjoy a scaling property with constant smoothness parameter and their distribution is close to Gaussian, so we model:
log(sig_{t+delta} - log(sig_t) = sig (W_{t+delta}^h - W_t^H)

- ⇒ Let $L^{H,\nu}$ be the law on $[0,1]$ of the process $e^{\nu W_t^H}$. Then the law of the volatility process on $[0,T]$ renormalized on $[0,1]$ $\frac{\sigma_{tT}}{\sigma_0}$ is $L^{H,\nu T^H}$.
- ⇒ If one observes the volatility on T = 10 years (2500 days) instead of T = 1 day, the parameter νT^H defining the law of the volatility is only multiplied by $2500^H \sim 3$. Therefore, one observes quite the same properties on a very wide range of time scales. The roughness of the volatility process (H = 0.14) implies a multi-scaling behavior of the volatility.

We can predict the fractional Brownian motion:

$$\mathbb{E}[W_{t+\Delta}^H|\mathcal{F}_t] = \frac{\cos(H\pi)}{\pi}\Delta^{H+\frac{1}{2}}\int_{-\infty}^{t}\frac{W_s^H}{(t-s+\Delta)(t-s)^{H+1/2}}ds$$

We can use that formula for the log volatility:

$$\mathbb{E}[\log(\sigma_{t+\Delta}^2)|\mathcal{F}_t] \sim \frac{\cos(H\pi)}{\pi}\int_0^1\frac{\log(\sigma_{t-\Delta u}^2)}{(u+1)u^{H+1/2}}du$$

The idea is to look as far back in the past as far we want to predict the future. If trying to predict volatility one week ahead, one should essentially look at the volatility over the last week. Our prediction writes:

$$\mathbb{E}[\sigma_{t+\Delta}^2|\mathcal{F}_t] = \exp\left(\mathbb{E}[\log(\sigma_{t+\Delta}^2)|\mathcal{F}_t] + 2\nu^2 c\Delta^{2H}\right)$$

Summary: Volatility is rough.

Microstructural foundations of rough volatility

Hawkes Process: self-exciting point process with intensity λ_t

$$\lambda_t = \mu + \sum_{0<J_i<t}\phi(t-J_i) = \mu + \int_{(0,t)}\phi(t-s)dN_s$$

Where μ positive real number, ϕ regression kernel and J_i points of the process before time t.

- ⇒ Order flow (Number of trades) is modeled with Hawkes Process.
- ⇒ We know already that the order flow is essentially the same thing as the integrated volatility (variance) if the time scale is large enough:

$$N_t \sim \int_0^t \sigma_s^2 ds$$

Main reasons to use Hawkes:

- Natural and tractable extension of Poisson. Poisson portrays iid. processes and Hawkes autoregressive ones (depend linearly on its own precious terms and stochastic variable = volatility).
- Easily interpretable.

Stability condition: $||\phi||_{L^1} < 1$. Required to get a stationary intensity with finite first moment and to obtain classical ergodic properties.

- ⇒ Models how many orders are triggered in the k-th generation. $||\phi||_{L^1}$ directly triggered orders by an exogenous one, $||\phi||_{L^1}^2$ triggered orders by a triggered order by an exogenous one, …
- ⇒ Average size of a cluster

$$\sum_{k\geq 1}||\phi||_{L^1}^k = \frac{||\phi||_{L^1}}{1-||\phi||_{L^1}}$$

Average proportion of endogenously triggered events is:

$$\frac{\|\phi\|_{L^1}/(1-\|\phi\|_{L^1})}{1+\|\phi\|_{L^1}/(1-\|\phi\|_{L^1})} = \|\phi\|_{L^1}$$

Limiting Behavior

Aim:

- Study nearly unstable Hawkes-Process ($\|\phi\|_{L^1}$ close to 1).
- Memory effect: $\phi(x) \sim \frac{K}{x^{1+\alpha}}$ for $x \to \infty$ (due to metaorders splitting).

Details: 6, Page 41-49

Limit theorem

For $\alpha > 1/2$, the sequence of renormalized Hawkes processes converges to some process which is differentiable on $[0,1]$. Moreover, the law of its derivative V satisfies

$$V_t = F^{\alpha,\lambda}(t) + \frac{1}{\sqrt{\mu^* \lambda}} \int_0^t f^{\alpha,\lambda}(t-s)\sqrt{V_s}dB_s^1,$$

with B^1 a Brownian motion and

$$f^{\alpha,\lambda}(x) = \lambda x^{\alpha-1} E_{\alpha,\alpha}(-\lambda x^\alpha).$$

Rough Heston model

Using fractional integration, we easily get that the equation for V_t on the preceding slide is equivalent to

$$V_t = \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\lambda(1-V_s)ds + \frac{1}{\Gamma(\alpha)}\sqrt{\frac{\lambda}{\mu^*}}\int_0^t (t-s)^{\alpha-1}\sqrt{V_s}dB_s.$$

Now recall Mandelbrot-van-Ness representation :

$$W_t^H = \int_0^t \frac{dW_s}{(t-s)^{\frac{1}{2}-H}} + \int_{-\infty}^0 \left(\frac{1}{(t-s)^{\frac{1}{2}-H}} - \frac{1}{(-s)^{\frac{1}{2}-H}}\right) dW_s.$$

Therefore we have a rough Heston model with $H = \alpha - 1/2$. Furthermore, for any $\varepsilon > 0$, Y has Hölder regularity $\alpha - 1/2 - \varepsilon$.

Microstructural Foundations

- It is clearly established that there is a linear relationship between cumulated order flow and integrated variance.
  $\Rightarrow$ *Derivative* of the order flow corresponds to spot variance.
- Endogeneity of the market + order splitting leads to superposition effect, which explains why the volatility is rough.
- Near instability together with a tail index $\alpha = 0.6$ correspond to $H \sim 0.1$.
- In fact, one can show that rough volatility is just a consequence of the no statistical arbitrage principle.