Advanced Empirical Finance: Topics and Data Science

Stefan Voigt

Spring 2024

University of Copenhagen and Danish Finance Institute

High-frequency econometrics

High-frequency trading (?)

High-frequency trading

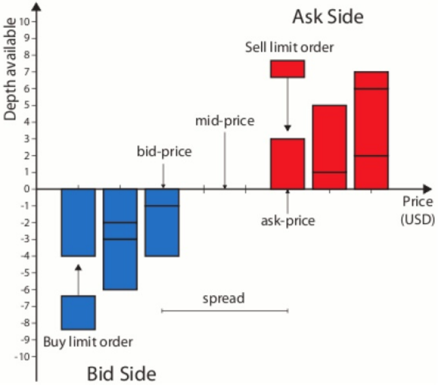
- So far, we have focused on quarterly, monthly, or daily price observations
- But: There is much more under the surface!
- High-frequency trading is a type of algorithmic financial trading characterized by high speeds, high turnover rates, and high order-to-trade ratios that leverages high-frequency financial data and electronic trading tools

Relevance for asset pricing

1. liquidity (transaction costs)
2. price informativeness (have you asked yourself how information makes it into prices?) (see also KU's market microstructure course)
3. more information may be beneficial for volatility estimation
4. wasteful investments?

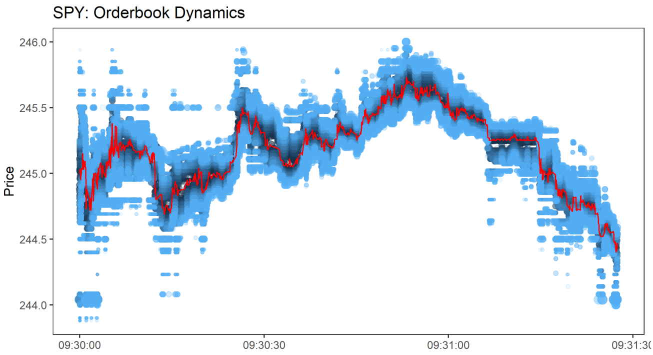
Quick primer on equity microstructure

- Market structure across asset classes can vary tremendously
- Large heterogeneity across US equity markets (lit pools, dark pools, maker-taker, ...)
- Standard framework: Continuous limit order book
- Limit orders indicate a willingness to buy/sell at a pre-specified price
- Market orders execute against limit orders



A typical NASDAQ trading day

- **Opening auction** at 9:30 am
- Continuous trading and order book adjustments throughout the entire day
- Message types: Cancellations/Submissions/Executions/ Hidden or Lit orders



- The number of daily messages is easily in the millions
- **Closing auction** at 4:00 pm

Realized volatility

HFT: Approaching continuous time finance

- Does it help to use data sampled at a high frequency to estimate μ_t and σ_t^2?
- Continuous-time finance helps
- Well-known example: Geometric Brownian motion (stock price movements are additive on the log scale)

$$\log(P_t) = X_t = X_0 + \mu t + \sigma W_t$$

where W_t is a Brownian motion

Brownian motion

- The process $(W_t)_{0 \leq t \leq T}$ is a Brownian motion provided that
  1. $W_0 = 0$
  2. $t \to W_t$ is a continuous function of t
  3. W has independent increments and for $t > s$, $W_t - W_s$ is normal with mean zero and variance $t - s$

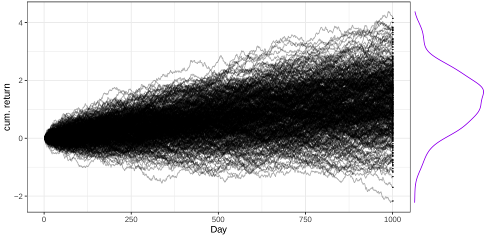
Brownian motion

- Suppose $t = 0$ is the start of the training day and $t = 1$ is the end of the day
- Assume there are n equidistant observations (transactions) of the log price
- One observation every $\Delta_{t_n} = 1/n$ units of time
- We observe $X_{t_{n,i}}$ with $t_{n,i} = i\Delta t_n$ and get

$$\Delta X_{t_{n,i}} = X_{t_{n,i}} - X_{t_{n,i-1}}$$

- By definition of the BM, the $\Delta X_{t_{n,i}}$ are iid normal $N(\mu\Delta t_n, \sigma^2\Delta t_n)$

Simulation of a Brownian motion



Estimating μ and σ^2 in the GBM model

- The natural estimator for $\hat{\mu}$ is

$$\hat{\mu}_n = \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} \Delta X_{t_n,i+1} = (X_1 - X_0)$$

- (Maybe) surprising: $\hat{\mu}_n$ does not depend on the sampling frequency. Consistent estimation requires $t \to \infty$
- (Maybe) more surprising: $\hat{\sigma}^2$ *can* be estimated consistently as $n \to \infty$!
- Set $U_{n,i} = \Delta X_{t_n,i} / (\sigma \Delta t_n^{1/2}) \sim N\left(\frac{\mu}{\sigma} \Delta t_n^{1/2}, 1\right)$ and define $\bar{U}_n = \frac{1}{n} \sum_{i=0}^{n-1} \left(U_{n,i} - \bar{U}_n\right)^2$
- Then

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{(n-1)\Delta t_n} \sum_{i=0}^{n-1} \left(\Delta X_{t_n,i+1} - \Delta \bar{X}_{t_n}\right)^2 \\
&= \frac{\sigma^2 \Delta t_n}{(n-1)\Delta t_n} \sum_{i=0}^{n-1} \left(\Delta U_{n,i} - \bar{U}_n\right)^2 \stackrel{L}{=} \sigma^2 \frac{\chi_{n-1}^2}{n-1}
\end{aligned}$$

- It follows that $E\left(\hat{\sigma}_n^2\right) = \sigma^2$ and $\mathrm{Var}\left(\hat{\sigma}_n^2\right) = \frac{2\sigma^4}{n-1}$

Non-centered estimator

- for high frequency data, the mean $\Delta\bar{X}_{t_n}$ is often **not** removed in estimation
- Instead, consider

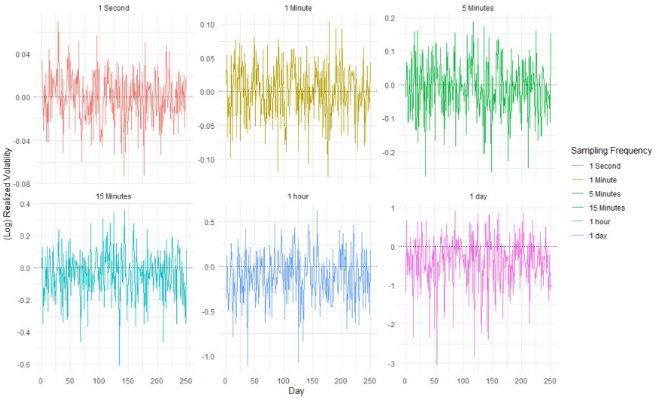
$$\hat{\sigma}^2_{n,\text{nocenter}} = \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} \left(\Delta X_{t_{n,i+1}}\right)^2 = \hat{\sigma}^2_n + \Delta t_n \bar{\mu}^2_n = \hat{\sigma}^2_n + \frac{1}{n}\bar{\mu}^2_n$$

- Since $\bar{\mu}_n$ does not depend on n, it follows that $\hat{\sigma}^2_{n,\text{nocenter}}$ is also consistent

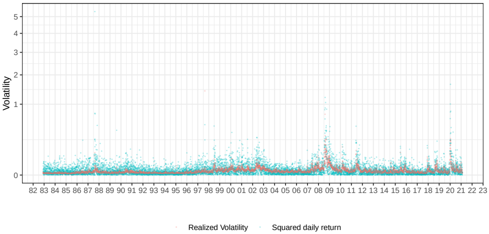
The realized volatility

- If μ_t and $\sigma_t > 0$ are predictable processes of finite variation and we assume the diffusion
$$dX_t = \mu_t dt + \sigma_t dW_t$$
the continuously compounded return over the time interval from $t - k$ to t is
$$r(t,k) = X_t - X_{t-k} = \int_{t-k}^{t} \mu_t dt + \int_{t-k}^{t} \sigma_t dW_t$$
- The diffusive sample path variation is called the integrated variance
$$IV(t-k,k) = \int_{t-k}^{t} \sigma_t^2 dt$$
- As long as there are no jumps, $IV(0,1) = RV^{(n)} := \sum_{i=0}^{n-1} r_{j,n}^2$
- $RV^{(n)}$ is the **realized variance**
- 'Infill' asymptotics: Sampling on the highest possible frequencies crucial!

Infill asymptotics



S&P 500 - Realized Volatility from 1988

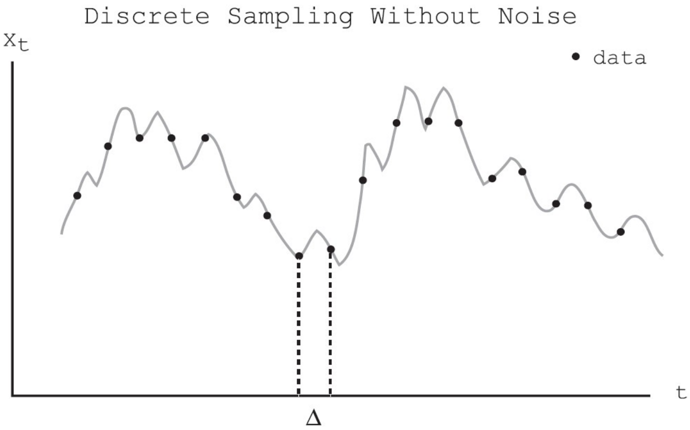


Market Microstructure Noise

Market Microstructure Noise

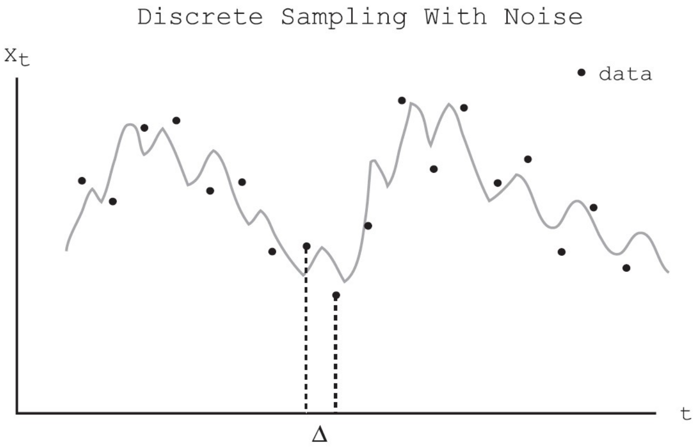
- Problem in practice: market microstructure frictions
- E.g., Bid-ask spreads, price discreteness, asymmetric information, strategic order placement
- We can only observe
  $$Y_{i\Delta n} = X_{i\Delta n} + U_{i\Delta n}$$
  where $Y_{i\Delta n}$ is the observed (log) transaction price or quote, $X_{i\Delta n}$ is the *efficient* latent (log) price
- $U_{i\Delta n}$ is white noise which captures microstructure frictions

What is the problem with microstructure noise?



- So far: the highest possible sampling frequency is optimal

What is the problem with microstructure noise?



- Now: What is the RV estimator’s optimal sampling frequency/adjustment in the presence of microstructure noise?

What is the problem with microstructure noise?

- Consider for now the easiest case with $X_t = \sigma W_t$ and $\tilde{X}_t = X_t + U_t$ where the U_t are iid noise with zero mean and variance a^2
- Then

$$
\begin{aligned}
Y_{\tau_i} &= \tilde{X}_{\tau_i} - \tilde{X}_{\tau_{i-1}} \\
&= \left(X_{\tau_i} - X_{\tau_{i-1}}\right) + U_{\tau_i} - U_{\tau_{i-1}} \\
&= \sigma\left(W_{\tau_i} - W_{\tau_{i-1}}\right) + U_{\tau_i} - U_{\tau_{i-1}}
\end{aligned}
$$

- We get $\operatorname{Var}\left(Y_{\tau_i}\right) = \sigma^2\Delta_n + 2\alpha^2$ and $\operatorname{Cov}\left(Y_{\tau_i}, Y_{\tau_{i-1}}\right) = -\alpha^2$
- the proportion of the total return variance that is market microstructure-induced is

$$
\pi = \frac{2\alpha^2}{2\alpha^2 + \sigma^2\Delta_n}
$$

- As Δ_n gets small, a larger portion of the volatility reflects noise!
- Noise bias adjustment possible (Ait-Sahalia et al., 2005)

Optimal sampling frequencies (Ait-Sahalia et al., 2005)

Optimal sampling frequency

	T		
Value of a	1 day	1 year	5 years
	(a) $\sigma = 30\%$ Stocks		
0.01%	1 min	4 min	6 min
0.05%	5 min	31 min	53 min
0.1%	12 min	1.3 h	2.2 h
0.15%	22 min	2.2 h	3.8 h
0.2%	32 min	3.3 h	5.6 h
0.3%	57 min	5.6 h	1.5 day
0.4%	1.4 h	1.3 day	2.2 days
0.5%	2 h	1.7 day	2.9 days
0.6%	2.6 h	2.2 days	3.7 days
0.7%	3.3 h	2.7 days	4.6 days
0.8%	4.1 h	3.2 days	1.1 week
0.9%	4.9 h	3.8 days	1.3 week
1.0%	5.9 h	4.3 days	1.5 week

- the presence of market microstructure noise makes it optimal to sample less often than would otherwise be the case in the absence of noise