PRINCETON UNIVERSITY

ORF 474: High Frequency Trading
Spring 2020

Robert Almgren

Lecture 8a

March 30, 2020

Today

- High frequency volatility
- What is volatility (Brownian motion)
- Effects of microstructure noise
- How to remove microstructure noise
- How to test volatility model?

  Simulated data

  Actual data
- Correlation
- Cointegration and price prediction (next few weeks)

Fundamental problem: serial correlation

prices $\quad p_0, p_1, \ldots, p_n$

suppose equal time spacing (for now)

$$p_j \;=\; p_{j-1} \;+\; \xi_j, \qquad \mathbb{E}(\xi_j) = 0,\; \mathbb{E}(\xi_j^2) = \sigma_1^2$$

$\sigma_1 \;=\;$ one-period volatility

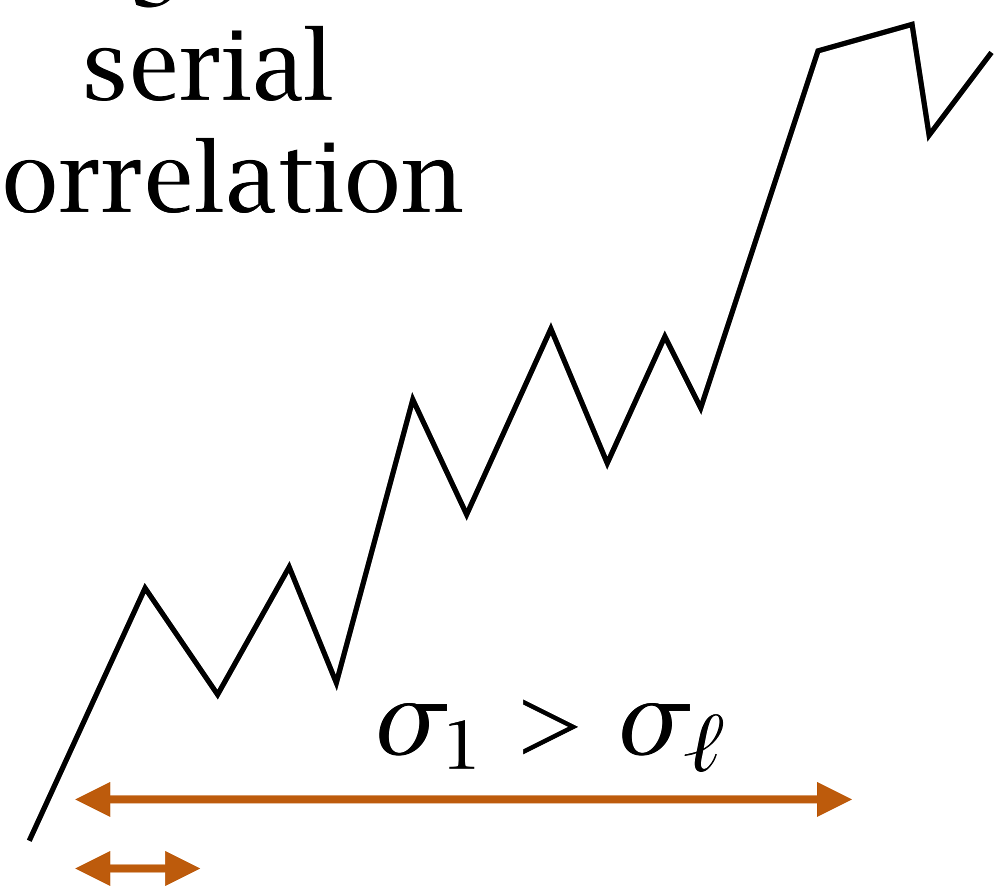
What is $\quad \sigma_\ell^2 \;=\; \dfrac{1}{\ell}\,\mathbb{E}\Big(\,(p_{k+\ell} - p_k)^2\,\Big) \quad ?$

does not depend on k if process is stationary

does not depend on ℓ if increments are independent

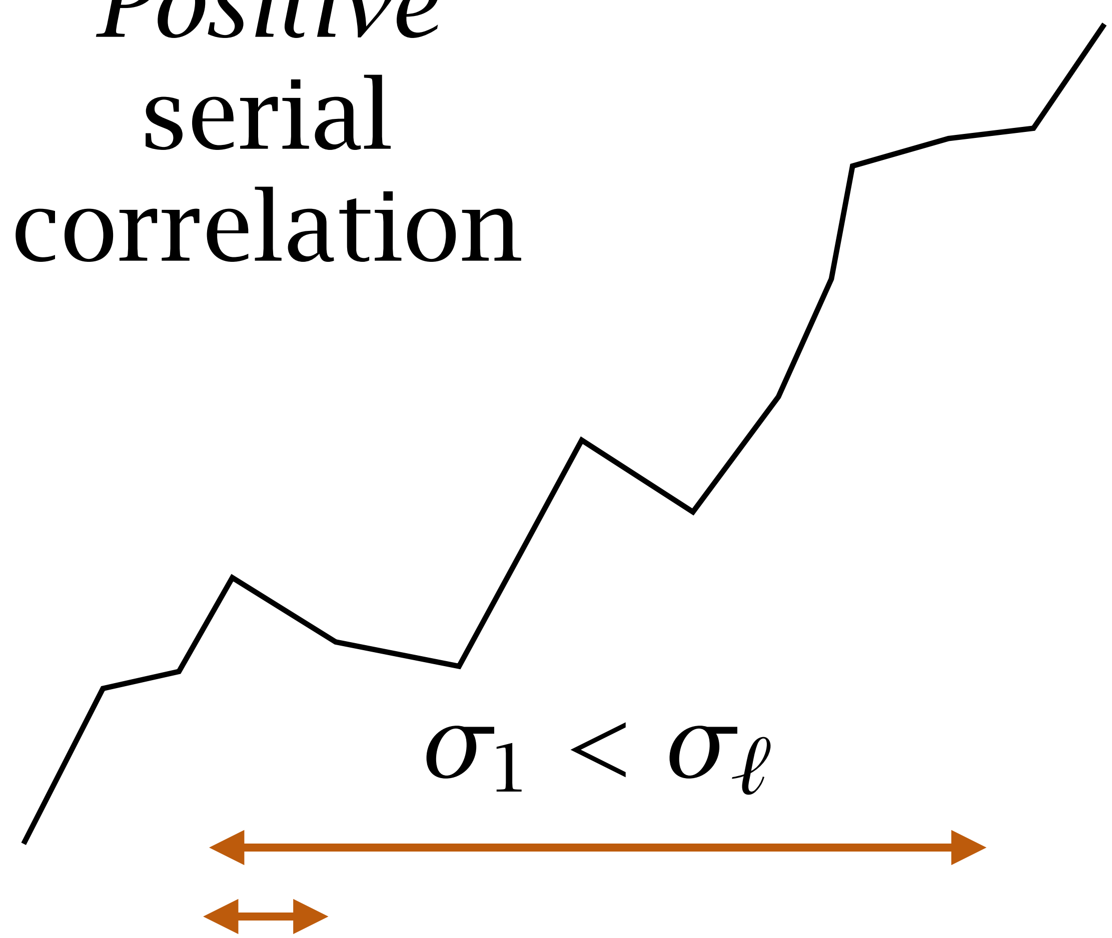
Intuition

Negative
serial
correlation



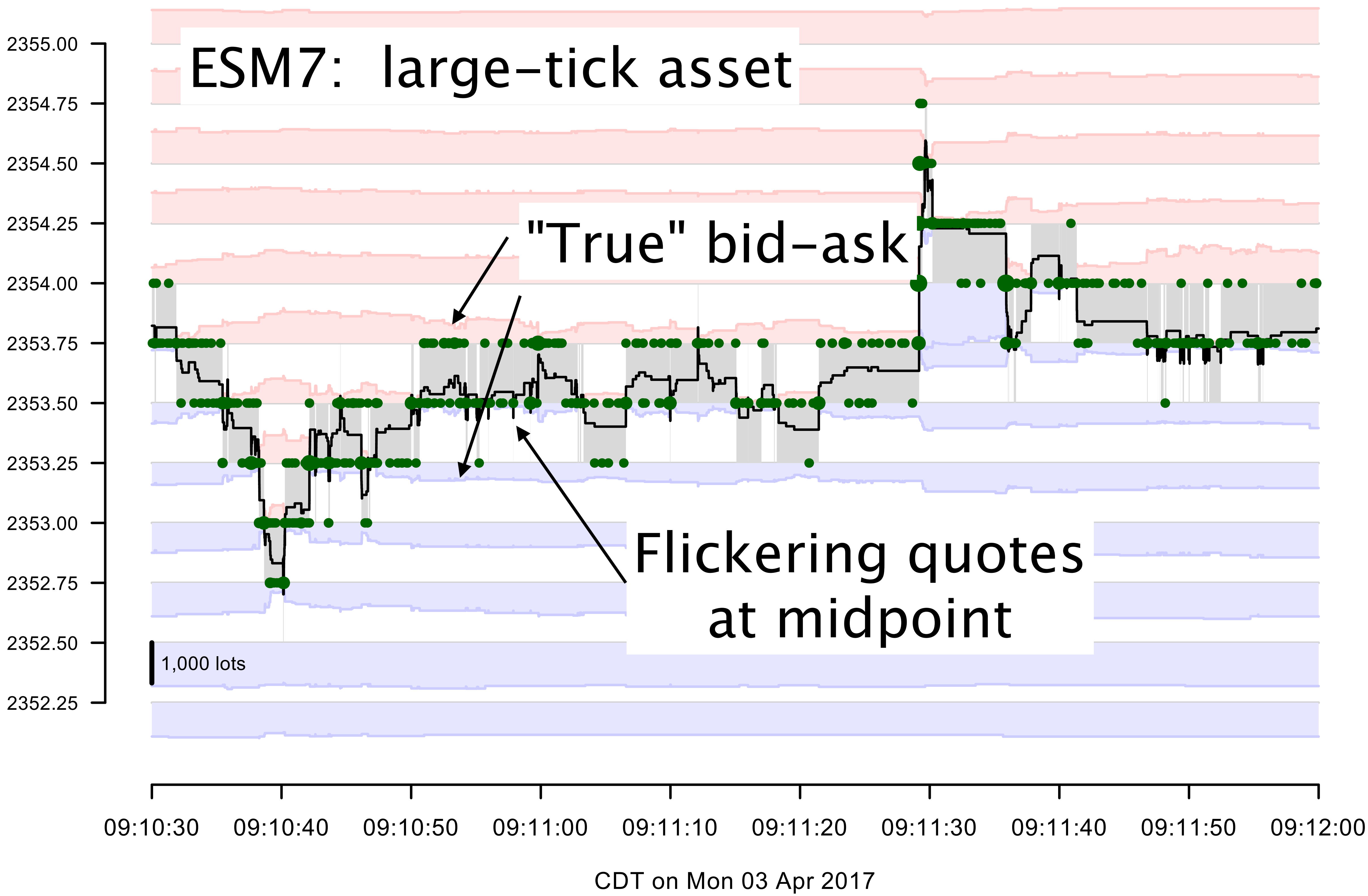
Local fluctuations
cancel
at longer scales

Positive
serial
correlation



Local fluctuations
reinforce
at longer scales

Prices do have negative serial correlation



Mean reversion

- Trade prices revert because of Roll costs
- Quote midpoints should be martingale
- In fact, quote midpoint motion is also reverting
- Volatility measurement is about correcting for this reversion

$$
\begin{aligned}
(p_{k+\ell} - p_k)^2 &= \left(\sum_{i=1}^{\ell} \xi_{k+i} \right)^2 \\
&= \sum_{i=1}^{\ell} \xi_{k+i}^2 + \sum_{i \neq j} \xi_{k+i} \xi_{k+j}
\end{aligned}
$$

If the ξ_i are uncorrelated, then

$$
\mathbb{E}\Big((p_{k+\ell} - p_k)^2 \Big) = \sum_{i=1}^{\ell} \mathbb{E}(\xi_{k+i}^2) = \ell\, \sigma_1^2
$$

$\sigma_\ell = \sigma_1$ Volatility independent of time scale

Suppose successive changes are correlated

$$\operatorname{corr}(\xi_i, \xi_{i+1}) = \rho$$

and further

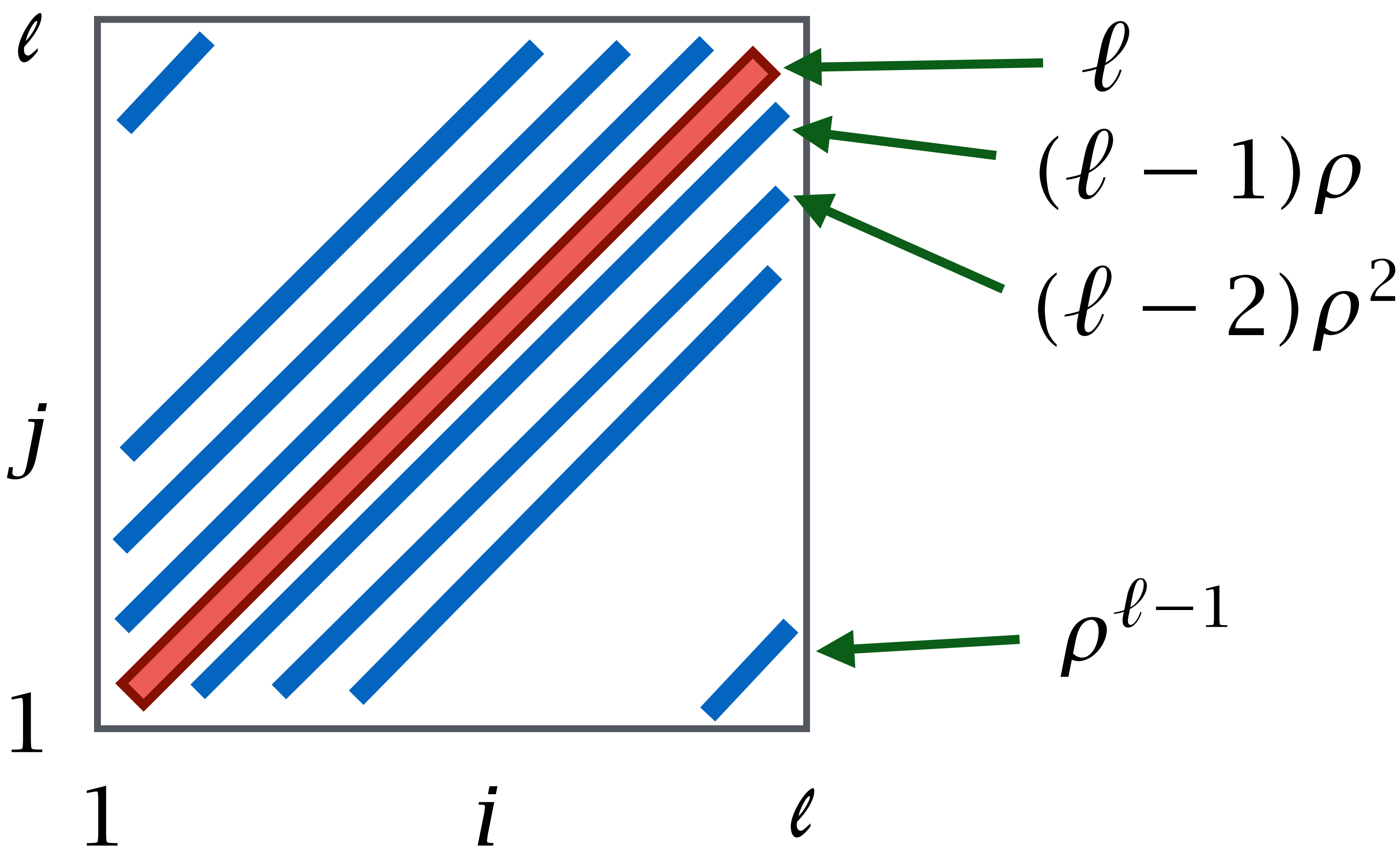
$$\operatorname{corr}(\xi_i, \xi_j) = \rho^{|i-j|}$$

(This is a very specific simple model and not fully descriptive of real correlation structures)

$$\sigma_\ell^2 = \frac{1}{\ell}\,\mathbb{E}\Big((p_{k+\ell} - p_k)^2 \Big) = \sigma_1^2 \left(1 + \frac{1}{\ell} \sum_{i \neq j} \rho^{|i-j|} \right)$$

Sum of the square

$$\frac{1}{\ell}\sum_{i\neq j}\rho^{|i-j|} = \frac{2}{\ell}\sum_{i=1}^{\ell-1}(\ell-i)\,\rho^i$$



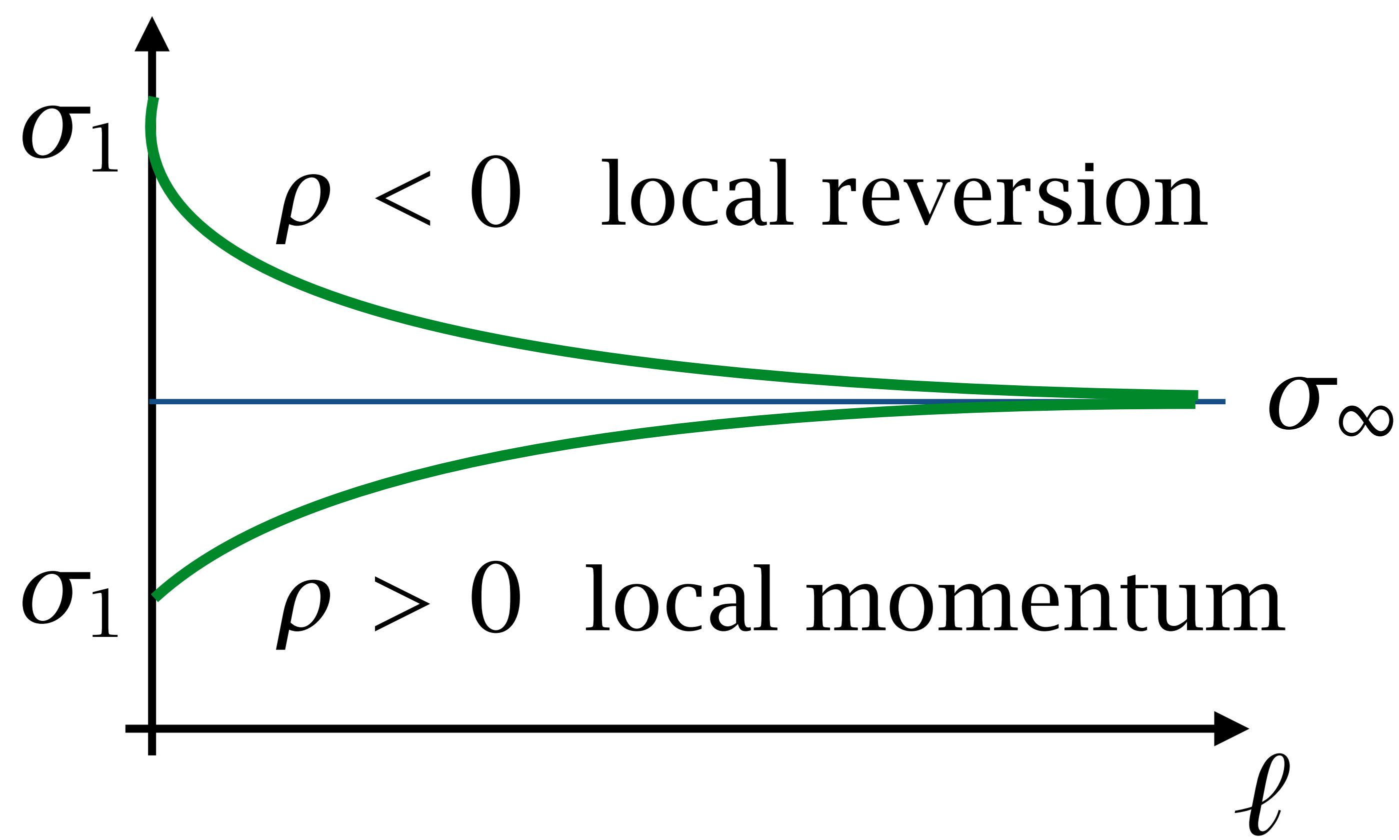
Long-lag limit

ℓ large

$$\frac{1}{\ell}\sum_{i\neq j}\rho^{|i-j|} = 2\sum_{i=1}^{\ell-1}\rho^i - \frac{2}{\ell}\sum_{i=1}^{\ell-1} i\,\rho^i$$

$$\approx 2\sum_{i=1}^{\infty}\rho^i = \frac{2\rho}{1-\rho}$$

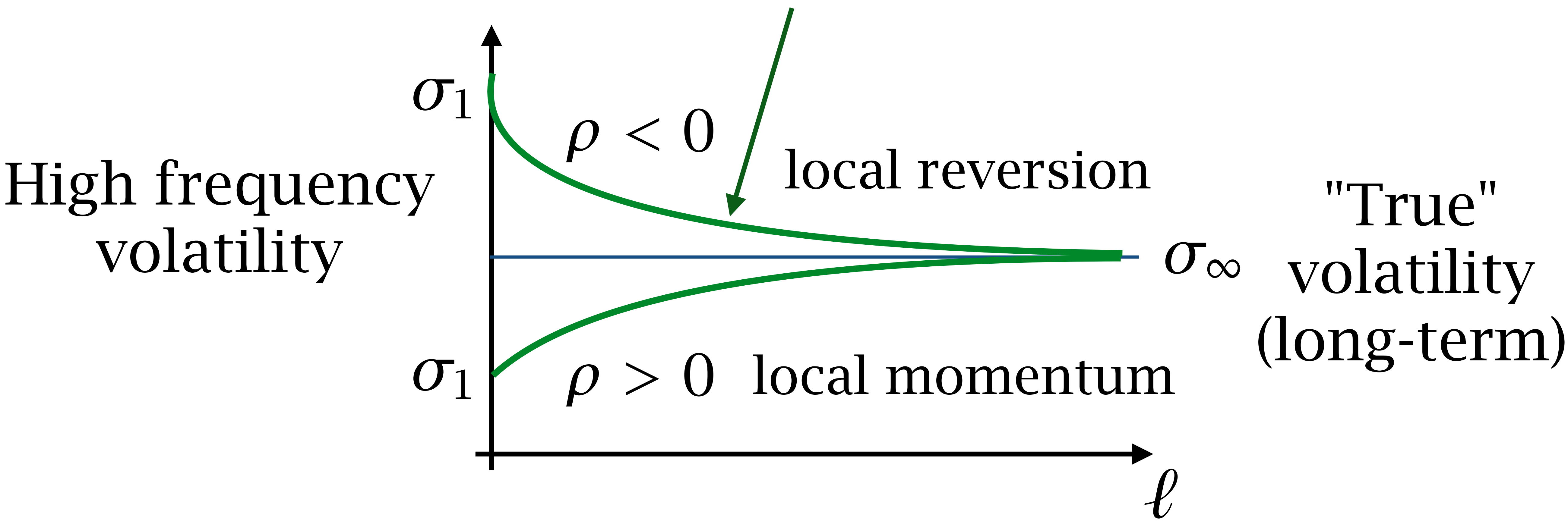
$$\sigma_\infty^2 = \frac{1+\rho}{1-\rho}\,\sigma_1^2$$



General lag

$$\sum_{i=1}^{\ell-1} (\ell - i)\rho^i \;=\; \rho\,\frac{(\ell-1) - \ell\rho + \rho^\ell}{(1-\rho)^2}$$

$$\sigma_\ell^2 \;=\; \sigma_\infty^2 \left(1 - \frac{2\rho}{\ell}\,\frac{1-\rho^\ell}{1-\rho^2} \right)$$



Volatility and variance

Price (or log-price) $dX(t) = \sigma(t)\, dB(t)$ no drift

Variance on given time interval

$$Q = \int_0^T \sigma(t)^2\, dt \qquad \Big(\approx \sigma^2 T \text{ if } \sigma(t) \text{ is constant}\Big).$$

Find estimator for Q given observations of X(t) at discrete times in [0,T].

No model for process **σ**(t), whether constant, deterministic, random, correlated with B(t) (leverage effect).
Only trying to measure historical value.
Forecasting is a different problem.

Want variance to be like traded volume

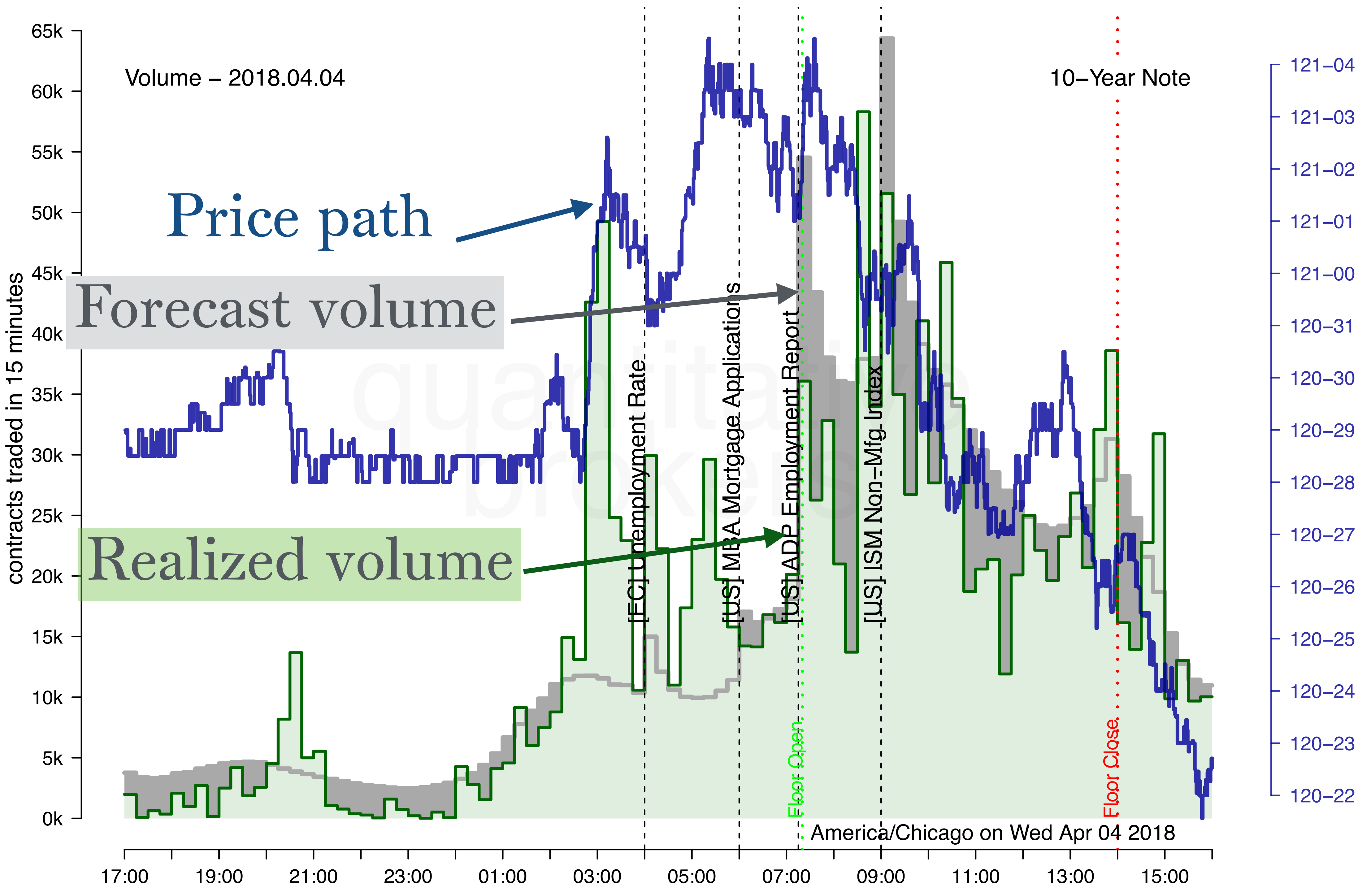
$$Q = \int_0^T \sigma(t)^2 \, dt, \quad \sigma(t)^2 = \text{price variance per unit time}$$

$$V = \int_0^T v(t) \, dt, \quad v(t) = \text{traded volume per unit time}$$

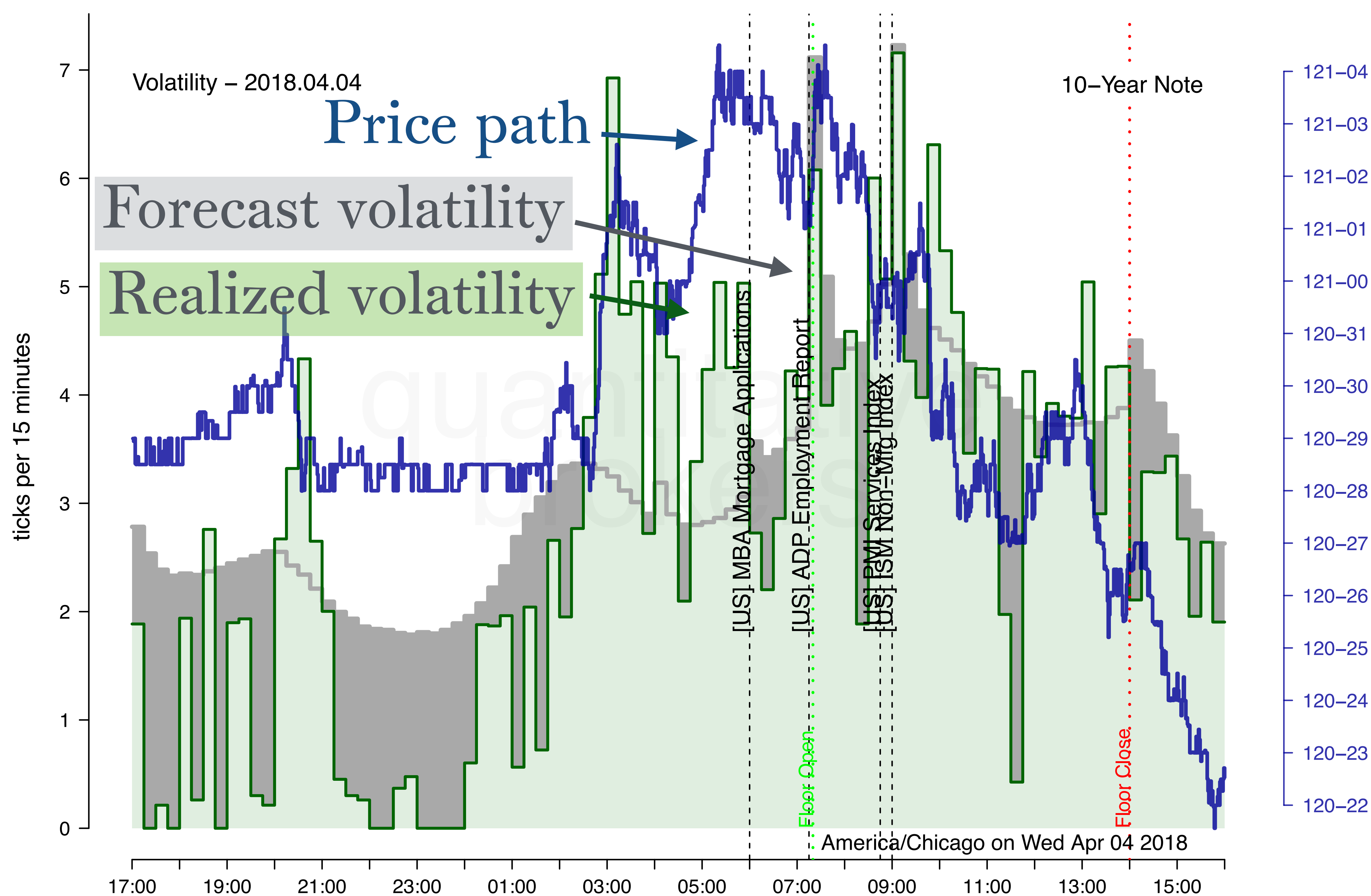
V is
- easy to measure by adding trades
- additive across different time intervals

Want to measure Q just as easily

Realized volume comparison with forecast



Realized variance comparison with forecast

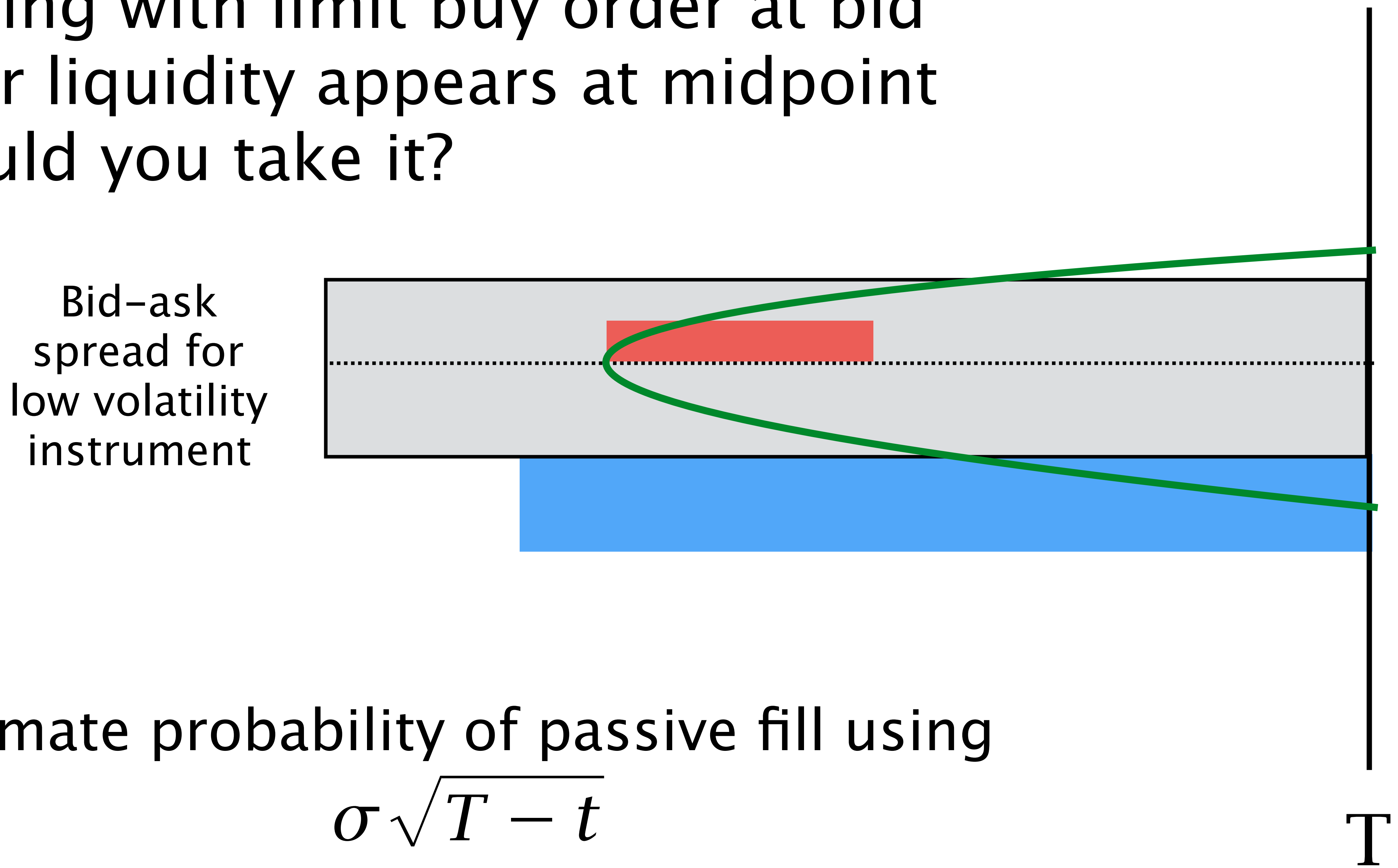


Why would you need volatility

- Descriptive interest
  - intraday structure
  - effect of events
- Estimated price motion over future time
  - Crossing decisions
  - Deviation from trade schedule

Crossing decisions

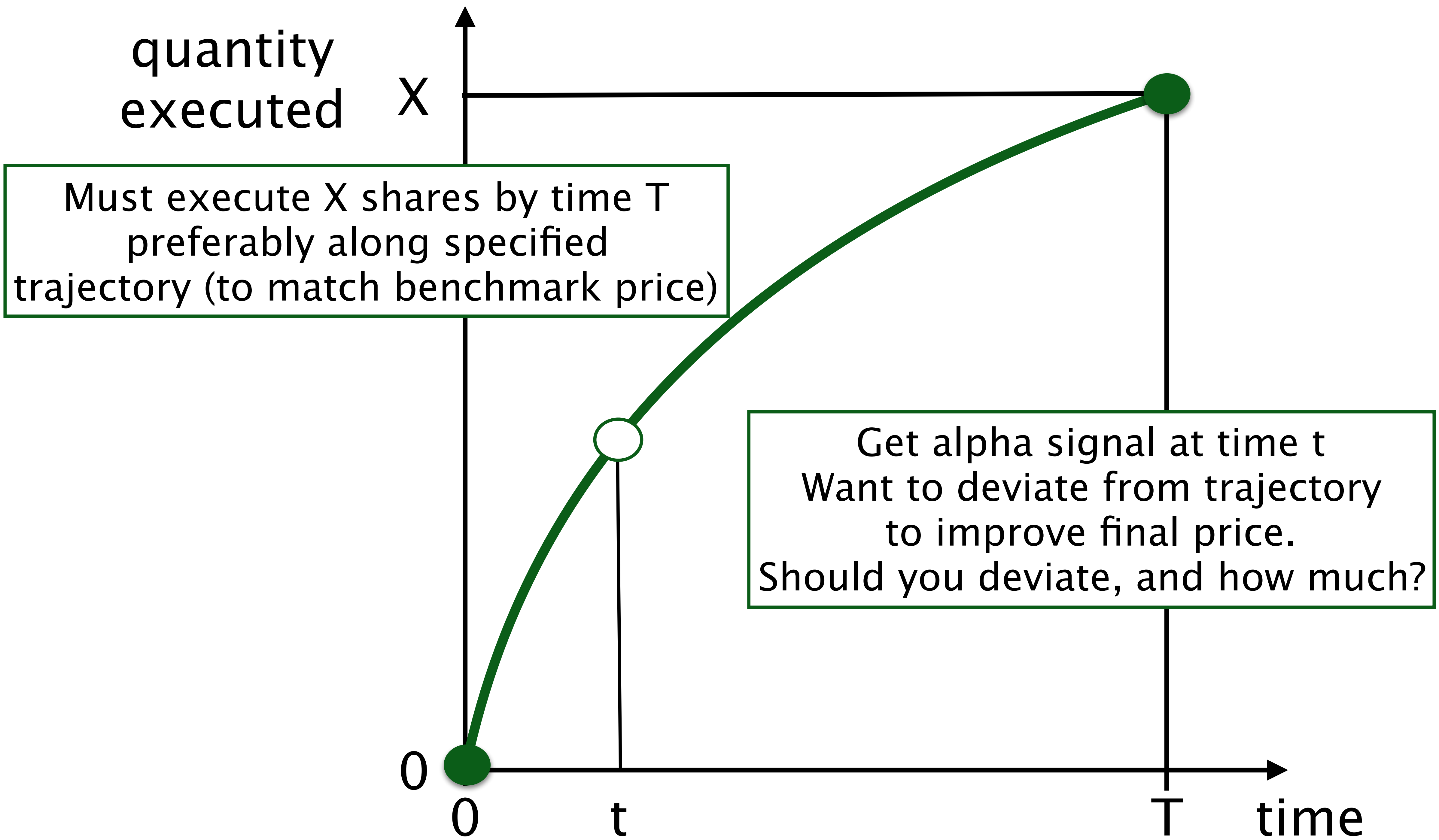
You must complete buy order by time T
Waiting with limit buy order at bid
Offer liquidity appears at midpoint
Should you take it?



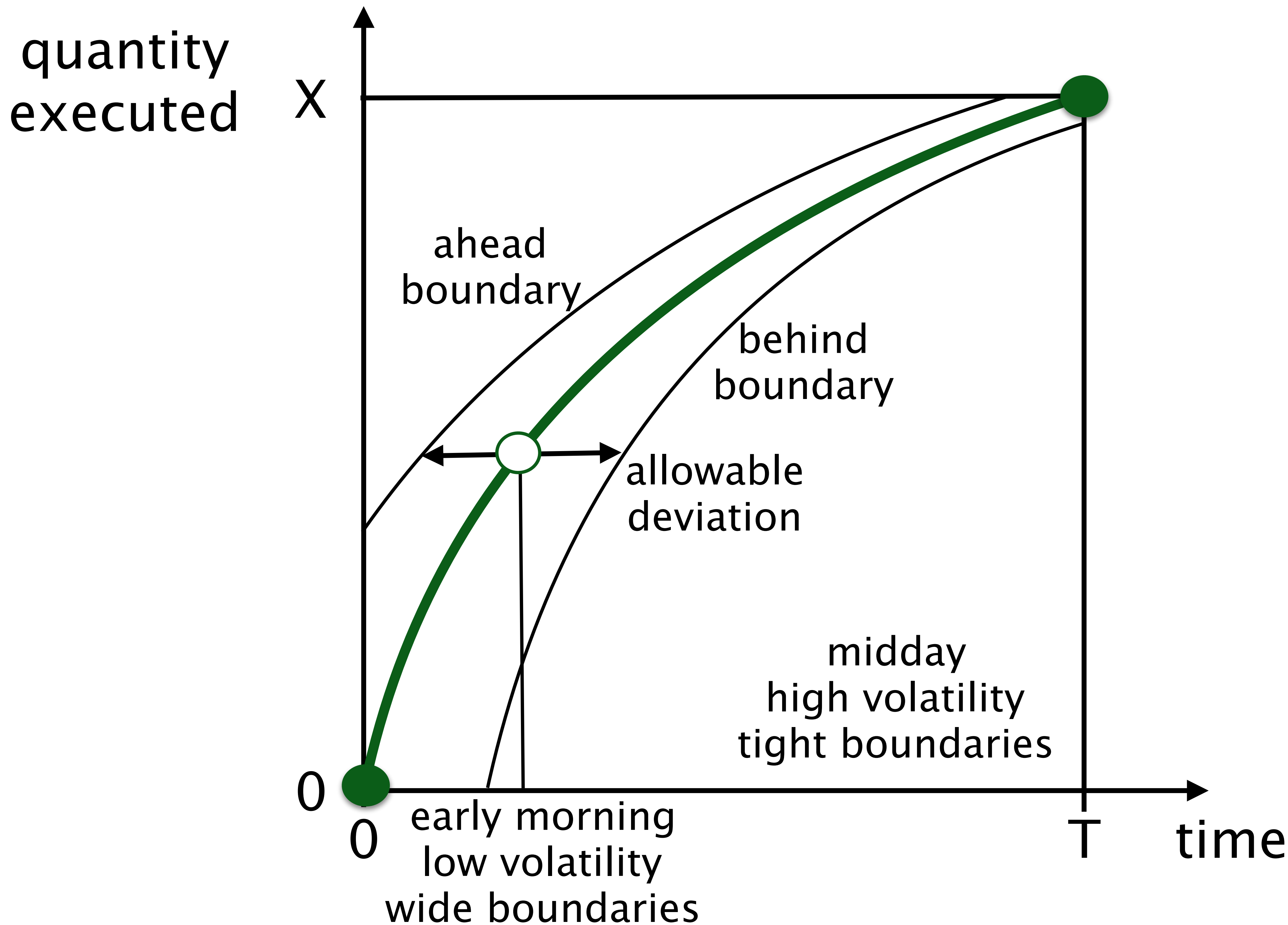
Estimate probability of passive fill using

$$\sigma\sqrt{T-t}$$

Deviation from trajectory



Draw "ahead" and "behind" boundaries
Deviations from trajectory determined by volatility risk, which varies through day.

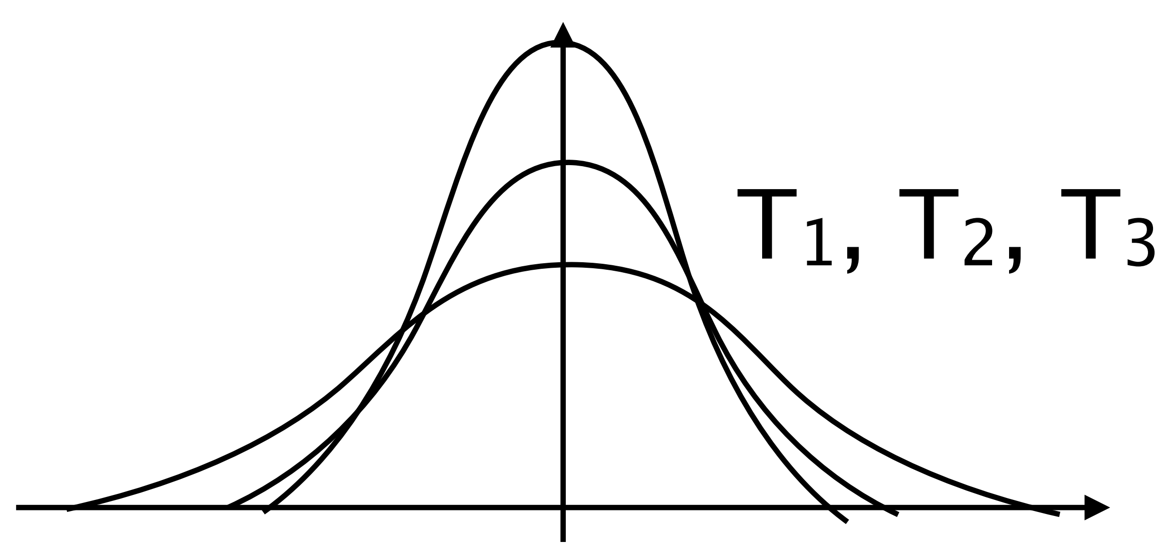


Critique of volatility

- Use forecast $\boldsymbol{\sigma}(t)$ for predicting size of price changes across forward time intervals

$$\Delta P(T) \sim \mathcal{N}(0, Q), \quad Q = \int_0^T \sigma(t)^2 \, dt$$

- Could also sample price changes directly rather than squeezing into parametric model



Realized vs implied volatility

- Realized: actual variation of price process
  - backward-looking (though can forecast)
  - most directly relevant for trading
- Implied: calculated from option prices (e.g. VIX)
  - forward-looking
  - model dependent
- Relation between them is interesting but complex

Estimate realized variance

$$Q \;=\; \int_0^T \sigma(t)^2\,dt \qquad \Big(\approx \sigma^2 T \text{ if } \sigma(t) \text{ is constant}\Big).$$

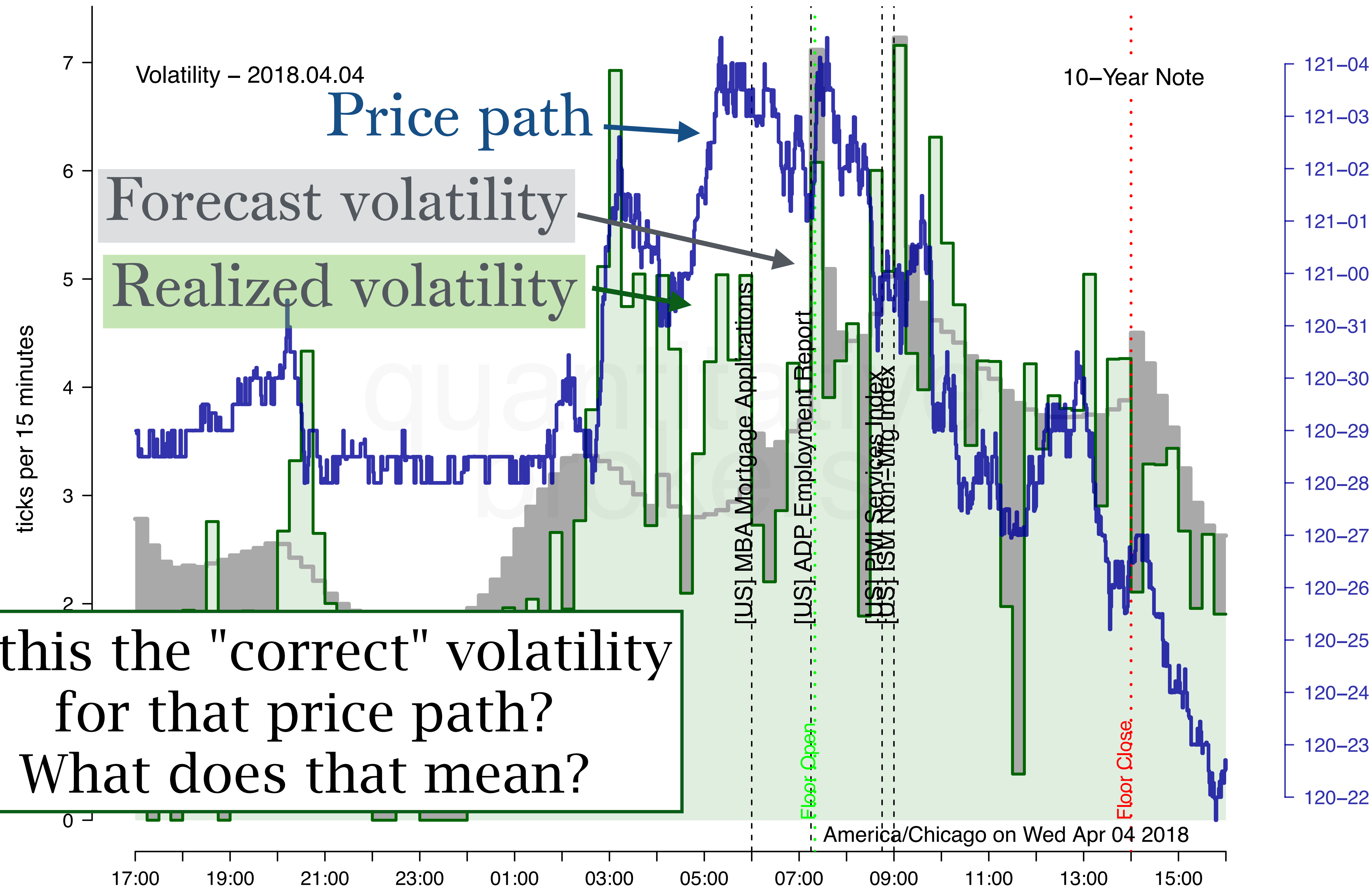
Given observations of price X(t)
generated from dX(t) = σ dB(t)

Uncertainty in estimators comes from
realizations of B(t) for given σ(t)

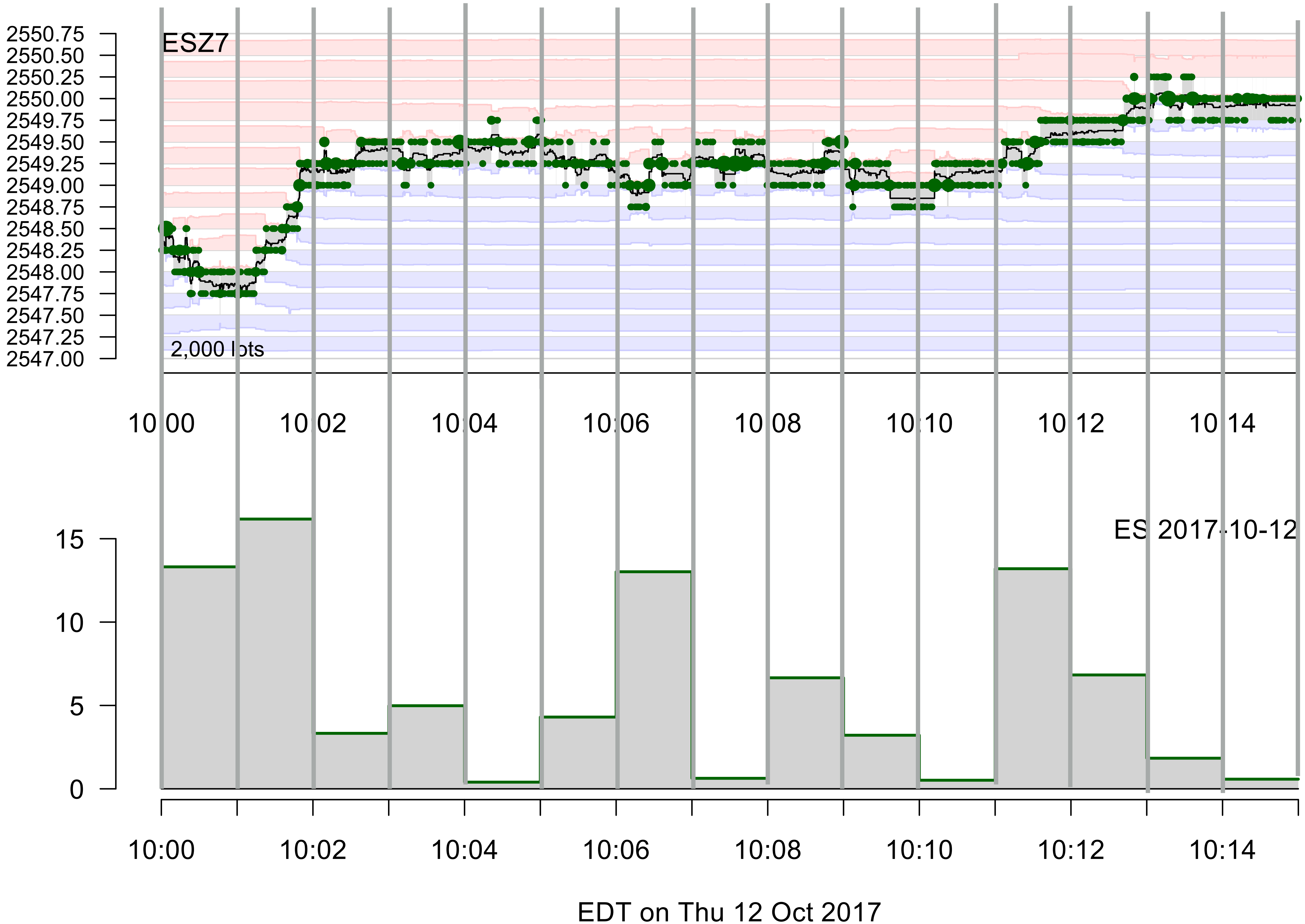
Challenges

- Accommodate microstructural noise
  to use all available data (tick-by-tick level)
- Difficult to know when it is "correct"
  real data is not generated from Brownian motion

Difficult to know when volatility is correct

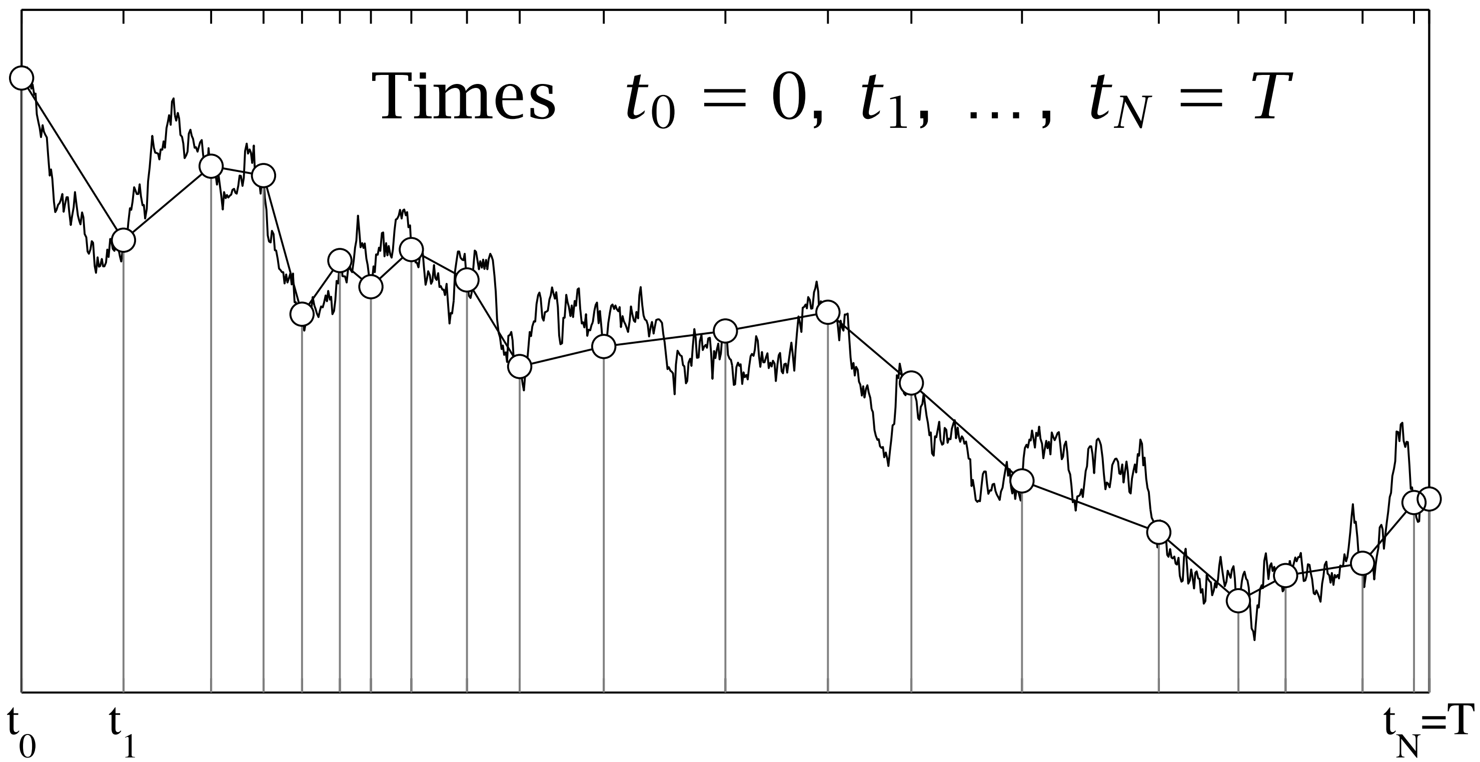


Is this the "correct" volatility
for that price path?
What does that mean?


ESZ7
2550.75
2550.50
2550.25
2550.00
2549.75
2549.50
2549.25
2549.00
2548.75
2548.50
2548.25
2548.00
2547.75
2547.50
2547.25
2547.00
2,000 lots
10:00
10:02
10:04
10:06
10:08
10:10
10:12
10:14
ES 2017-10-12
15
10
5
0
10:00
10:02
10:04
10:06
10:08
10:10
10:12
10:14
EDT on Thu 12 Oct 2017

Best case: no local noise

Time points non-random
or at least independent of X(t)



Exact samples

$$x_j \;=\; X(t_j)$$

$$x_j - x_{j-1} \;=\; \int_{t_{j-1}}^{t_j} \sigma(t)\, dB(t)$$

$x_j - x_{j-1}$ is normal

$$\mathbb{E}(x_j - x_{j-1}) = 0, \qquad \mathrm{Var}(x_j - x_{j-1}) = Q_j \equiv \int_{t_{j-1}}^{t_j} \sigma(t)^2\, dt.$$

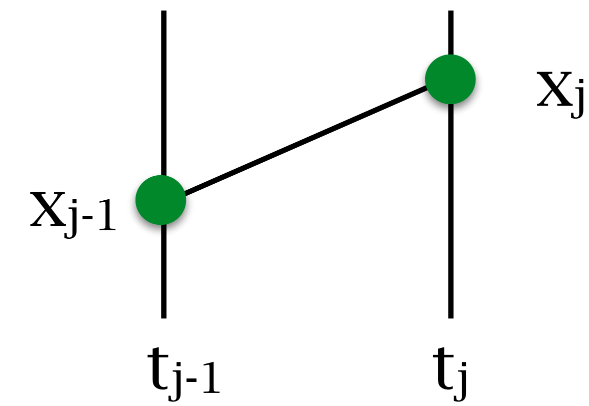
$(x_j - x_{j-1})^2$ is χ^2

$$\mathbb{E}\big((x_j - x_{j-1})^2\big) = Q_j, \qquad \mathrm{Var}\big((x_j - x_{j-1})^2\big) = 2Q_j^2.$$

First: σ constant number

Best estimator of σ on single interval

$$\hat{\sigma}_j^2 = \frac{(x_j - x_{j-1})^2}{t_j - t_{j-1}}$$



$$\mathbb{E}(\hat{\sigma}_j^2) = \sigma^2 \qquad \mathrm{Var}(\hat{\sigma}_j^2) = 2\sigma^4.$$

Variance of estimator is independent of Δt
Optimal combination weights all equally.

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{j=1}^{N} \frac{(x_j - x_{j-1})^2}{t_j - t_{j-1}}.$$

Division is always scary

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2 \qquad \mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{N}.$$

Unbiased; has smallest variance for given data.

σ(t) not constant

Realized variance: $$\mathrm{RV} \;=\; \sum_{j=1}^{N} (x_j - x_{j-1})^2 .$$

$$\mathbb{E}(\mathrm{RV}) = \sum_{j=1}^{N} Q_j = Q, \qquad \mathrm{Var}(\mathrm{RV}) = 2\sum_{j=1}^{N} Q_j^2. \qquad Q_j = \int_{t_{j-1}}^{t_j} \sigma(t)^2\, dt.$$

If N large, σ not far from constant, t_j reasonably spaced, then

$$\mathrm{RV} \;=\; Q \;+\; \xi, \quad \text{with } \mathrm{Var}(\xi) \sim \mathcal{O}\left(\frac{\sigma^4 T^2}{N}\right) \qquad Q = \int_0^T \sigma(t)^2\, dt.$$

RV is unbiased estimator of Q, error is smaller if N is larger

σ estimator from Realized Variance

$$\sigma_{\mathrm{RV}}^2 = \frac{\mathrm{RV}}{T} = \frac{1}{T}\sum_{j=1}^{N}(x_j - x_{j-1})^2$$

from before $$\hat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}\frac{(x_j - x_{j-1})^2}{t_j - t_{j-1}}.$$

equal if and only if all $t_j - t_{j-1}$ are same

Which is better? Should you divide each Δx^2 by Δt?

$$\frac{\mathrm{Var}(\sigma_{\mathrm{RV}}^2)}{\mathrm{Var}(\hat{\sigma}^2)} = \frac{N}{T^2}\sum_{j=1}^{N}(t_j - t_{j-1})^2 = \frac{\frac{1}{N}\sum(t_j - t_{j-1})^2}{\left(\frac{1}{N}\sum(t_j - t_{j-1})\right)^2} \geq 1$$

average of square of time intervals /
square of average

RV is less precise but more robust

Garman Klass approach

Mark B. Garman and Michael J. Klass
University of California, Berkeley

On the Estimation of Security Price Volatilities from Historical Data*

(*Journal of Business*, 1980, vol. 53, no. 1)

Improved estimators of security price volatilities are formulated. These estimators employ data of the type commonly found in the financial pages of a newspaper: the high, low, opening, and closing prices and the transaction volume. The new estimators are seen to have relative efficiencies that are considerably higher than the standard estimators.

High / low are not sample points that are independent of process realization

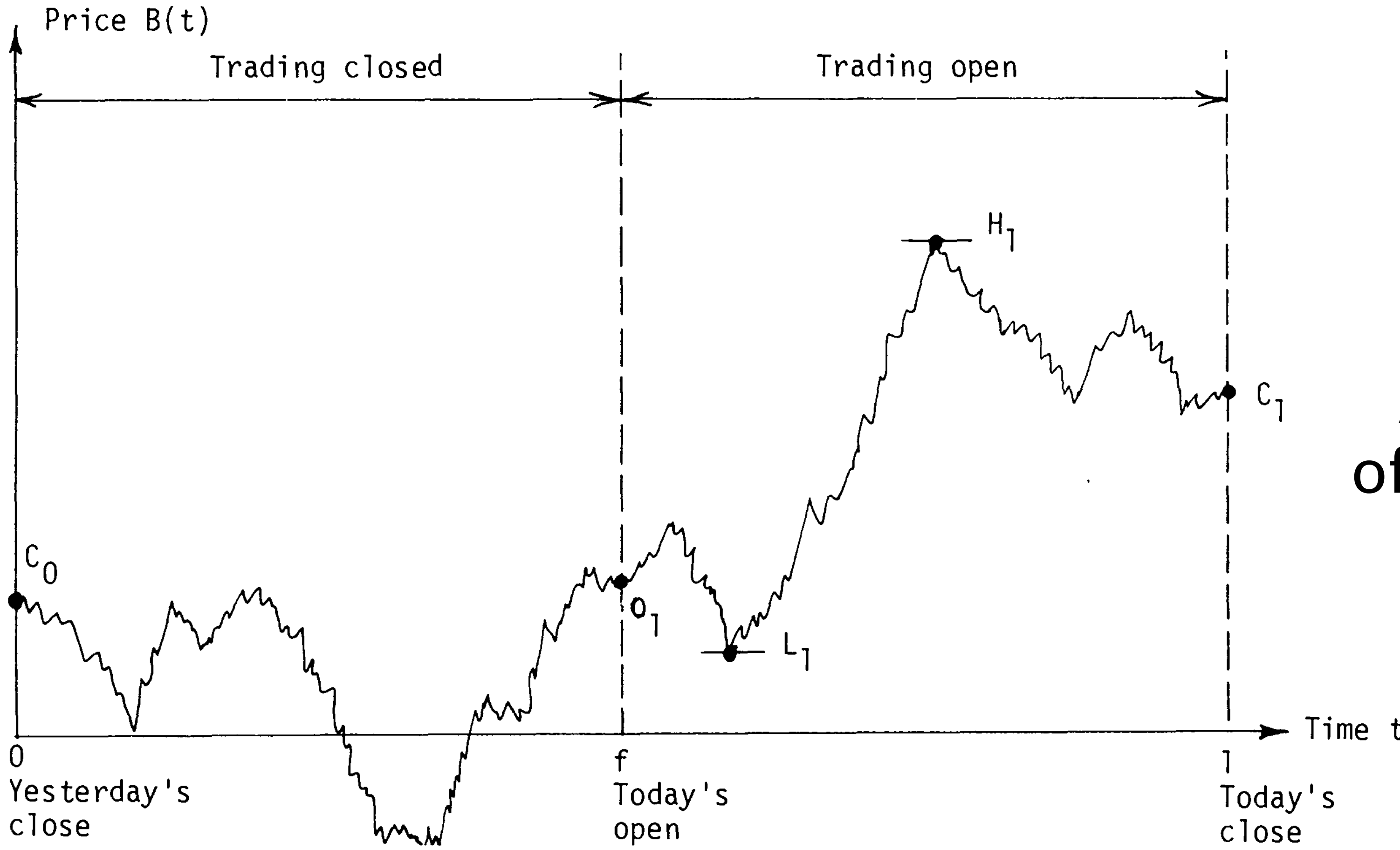


FIG. 1.—Price versus time

GK wanted daily volatility using open, close, high, low as reported in newspapers. Also wanted to take account of overnight no-trading period.

σ^2 = unknown variance (volatility) of price change;
f = fraction of the day (interval [0, 1]) that trading is closed;
$C_0 = B(0)$, previous closing price;
$O_1 = B(f)$, today's opening price;
$H_1 = \max_{f \leq t \leq 1} B(t)$, today's high;
$L_1 = \min_{f \leq t \leq 1} B(t)$, today's low;
$C_1 = B(1)$, today's close;
$u = H_1 - O_1$, the normalized high;
$d = L_1 - O_1$, the normalized low;
$c = C_1 - O_1$, the normalized close;
$g(u, d, c; \sigma^2)$ = the joint density of (u, d, c) given σ^2 and $f = 0$.

The classical estimator $\hat{\sigma}_0^2$ will provide the benchmark by which we shall judge all other estimators. Therefore, define the relative efficiency of an arbitrary estimator $\hat{y}$ by the ratio

$$\mathrm{Eff}(\hat{y}) \equiv \frac{\mathrm{var}(\hat{\sigma}_0^2)}{\mathrm{var}(\hat{y})}. \tag{4}$$

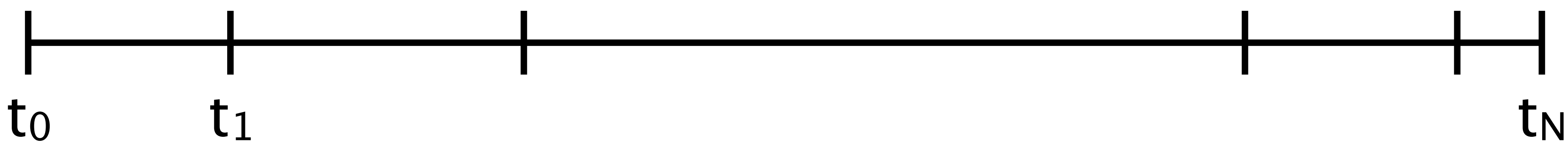
■ ■ ■

where $\zeta(3) = \sum_{k=1}^{\infty} 1/k^3 = 1.2021$ is Riemann's zeta function. Substituting the above moments into (17a) and (17b) via (18), we find that $a_1^* = 0.511$ and $a_2^* = -0.019$. Employing these values in (15) yields the best analytic scale-invariant estimator

$$\hat{\sigma}_4^2 \equiv 0.511(u - d)^2 - 0.019[c(u + d) - 2ud] - 0.383c^2. \tag{19}$$

We find that $\mathrm{Eff}(\hat{\sigma}_4^2) \approx 7.4$. (The more "practical" estimator $\hat{\sigma}_5^2 \equiv 0.5[u - d]^2 - [2 \log_e 2 - 1]c^2$ has virtually the same efficiency but eliminates the cross-product terms.)

in our notation, multiple time bins



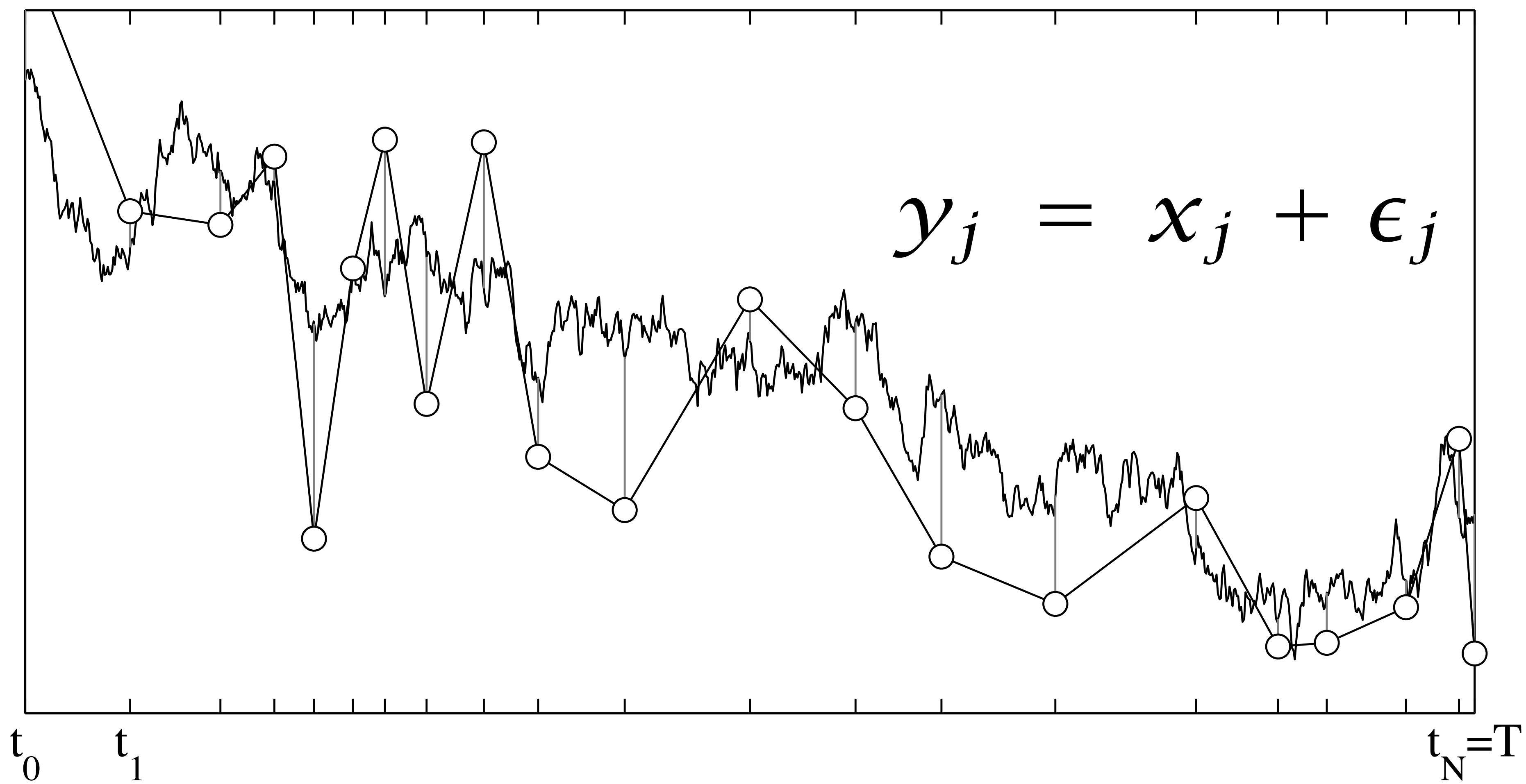
$$H_j = \sup_{t_{j-1} \le t \le t_j} X(t) \qquad L_j = \inf_{t_{j-1} \le t \le t_j} X(t)$$

(in addition to the $x_j = X(t_j)$)

$$\mathrm{RV}_{\mathrm{GK}} = \sum_{j=1}^{N} \left(\frac{1}{2} (H_j - L_j)^2 - (2 \log 2 - 1) \underbrace{(x_j - x_{j-1})^2}_{\text{as before}} \right).$$

Variance is 1/8 of variance based only on the x_j
(under assumptions of Brownian motion)
In practice: fairly robust and efficient

Microstructure noise



ϵ_j i.i.d. independent of $x(t)$, $\mathbb{E}(\epsilon_j) = 0$, $\mathbb{E}\left(\epsilon_j^2\right) = \eta^2$

ε_j represents microstructure (bid-ask, etc) noise per observation

Specific model: precise conclusions valid only for this model

Observed price changes have negative serial correlation

$$\Delta y_j = \Delta x_j + \Delta \epsilon_j$$

$$\begin{aligned}
\mathbb{E}(\Delta y_j \cdot \Delta y_{j+1}) \\
&= \mathbb{E}\big((\Delta x_j + \epsilon_j - \epsilon_{j-1})(\Delta x_{j-1} + \epsilon_{j+1} - \epsilon_j)\big) \\
&= \mathbb{E}(-\epsilon_j^2) \\
&= -\eta^2
\end{aligned}$$

Expect higher volatility at microstructure level

Journal of Econometrics 160 (2011) 160–175

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Ultra high frequency volatility estimation with dependent microstructure noise[☆]

Yacine Aït-Sahalia[a,*], Per A. Mykland[b], Lan Zhang[c]

[a] Princeton University and NBER, Department of Economics, Princeton, NJ 08544-1021, United States
[b] The University of Chicago, United States
[c] University of Illinois - Chicago, United States

ARTICLE INFO

Article history:
Available online 6 March 2010

Keywords:
Market microstructure
Serial dependence
High frequency data
Realized volatility
Subsampling
Two scales realized volatility
Multiple scales realized volatility

ABSTRACT

We analyze the impact of time series dependence in market microstructure noise on the properties of estimators of the integrated volatility of an asset price based on data sampled at frequencies high enough for that noise to be a dominant consideration. We show that combining two time scales for that purpose will work even when the noise exhibits time series dependence, analyze in that context a refinement of this approach is based on multiple time scales, and compare empirically our different estimators to the standard realized volatility.

1. Introduction

When studying financial data, the notion that noise plays an essential role is an accepted fact of life, whether at the high frequency typical of transactions data or at the lower frequencies more commonly used in asset pricing. The fact that this is a central issue is perhaps best demonstrated by the fact that two recent presidential addresses to the American Finance Association have been entitled "noise" (Black, 1986) and "frictions" (Stoll, 2000) respectively. So we work under the assumption that the observed log-price Y (either transaction or quoted) in high frequency financial data is the unobservable efficient log-price X, plus some noise component ϵ due to the imperfections of the trading process,

$$\boxed{Y_t = X_t + \epsilon_t.} \tag{1.1}$$

Since X is defined implicitly (as opposed to explicitly, such as the sum of expected discounted dividends for instance) we have maintained the simple identifying assumption that ϵ is independent of the X process. It is shown in Li and Mykland (2007) that this assumption can be substantially weakened (see also Jacod (1996) and Delattre and Jacod (1997)).

RV on observed prices

$$\begin{aligned}
\mathrm{RV}(y) &= \sum_{j=1}^{N} (y_j - y_{j-1})^2 = \sum_{j=1}^{N} \Big((x_j - x_{j-1}) + (\epsilon_j - \epsilon_{j-1}) \Big)^2 \\
&= \sum_{j-1}^{N} (x_j - x_{j-1})^2 + 2 \sum_{j-1}^{N} (x_j - x_{j-1})(\epsilon_j - \epsilon_{j-1}) \\
&\qquad + \sum_{j-1}^{N} (\epsilon_j - \epsilon_{j-1})^2 \\
&= Q + \underbrace{2N\eta^2}_{\text{bias}} + \xi
\end{aligned}$$

$$\mathrm{Var}(\xi) \sim \mathcal{O}\Big(N(\eta^4 + \sigma^4 \tau^2) \Big) \qquad \tau = T/N$$

Increasing N (finer sampling) increases both bias and variance (unless η=0)

Variance calculation

Term {1}

$$\sum_{j=1}^{N} (x_j - x_{j-1})^2 \;\sim\; Q \;+\; \xi_1, \qquad \mathrm{Var}(\xi_1) \;=\; \frac{\sigma^4 T^2}{N}$$

Term {2}

$$\sum_{j-1}^{N} (x_j - x_{j-1})(\epsilon_j - \epsilon_{j-1}) \quad \text{has mean 0}$$

$$\mathrm{Var} \sum_{j=1}^{N} (x_j - x_{j-1})(\epsilon_j - \epsilon_{j-1}) \;\approx\; \sum_{j=1}^{N} \mathrm{Var}(x_j - x_{j-1}) \cdot \mathrm{Var}(\epsilon_j - \epsilon_{j-1})$$

$$= \; \sigma^2 T \cdot 2\eta^2$$

Term {3}

$$(\epsilon_j - \epsilon_{j-1})^2 \;=\; \epsilon_j^2 + \epsilon_{j-1}^2 - 2\epsilon_j\epsilon_{j-1}$$

has mean $2\eta^2$, variance $\sim \eta^4$

$$\sum(\epsilon_j - \epsilon_{j-1})^2 \;=\; 2N\eta^2 \;+\; \xi_3 \qquad\qquad \mathrm{Var}(\xi_3) \approx N\eta^4$$

$$\mathrm{RV}(y) \;=\; \{1\} + \{2\} + \{3\} \;=\; Q \;+\; 2N\eta^2 \;+\; \xi$$

$$\mathrm{Var}(\xi) \;\sim\; N(\sigma^4\tau^2 + \eta^4) \;=\; \frac{\sigma^4 T^2}{N} \;+\; N\eta^4$$

Sampling error
decreases as N increases

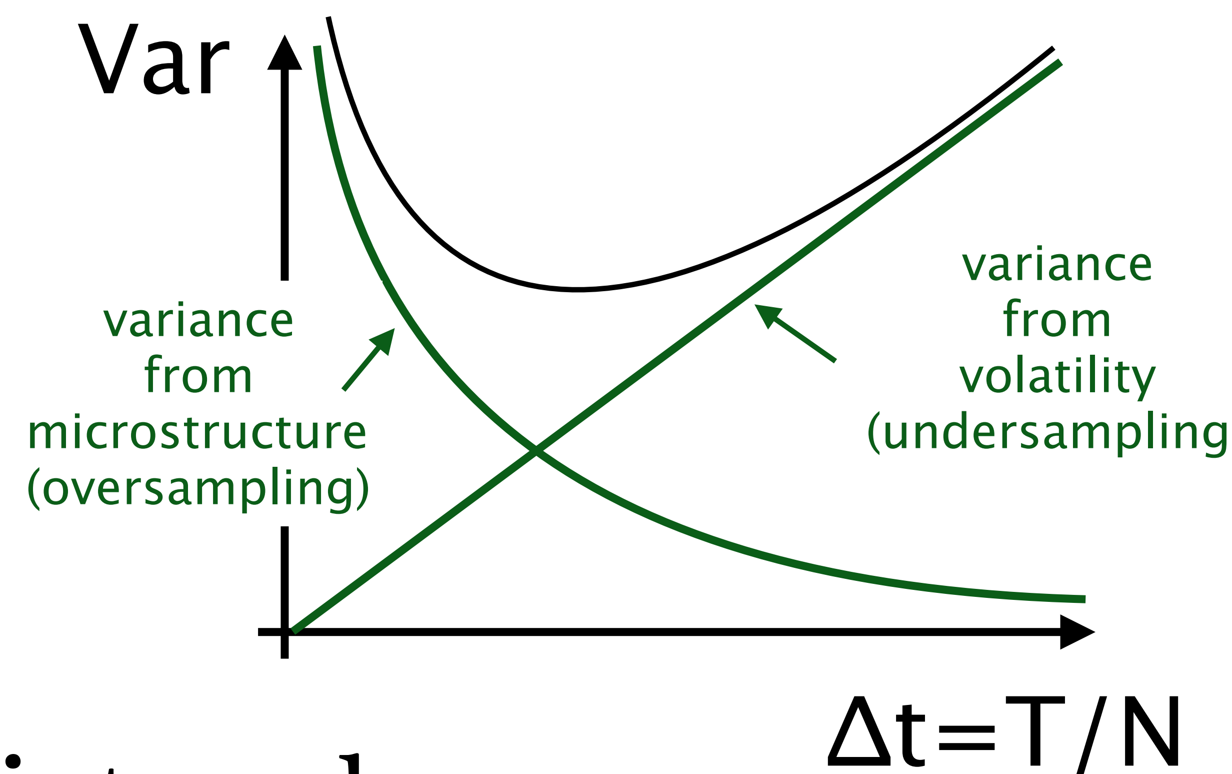
Error from microstructure
increases as N increases

Prevents you from using all your tick data

Optimal number of sampling points

$$\mathrm{Var}(\xi) \;\sim\; N(\sigma^4\tau^2 + \eta^4) \;=\; \frac{\sigma^4 T^2}{N} \;+\; N\eta^4$$

Minimize over N



$$N = \frac{\sigma^2 T}{\eta^2} \qquad \sigma^2\frac{T}{N} = \eta^2$$

Volatility on each time interval
~ same size as microstructure noise

Mitigation strategies

- Subsample (5-minutes, etc)
- Overlapping subsamples
- Two-scale correction (Zhang, Mykland, Ait-Sahalia)
- Improved sampling (Zhou)

Subsampling

How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise

Yacine Aït-Sahalia
Princeton University and NBER

Per A. Mykland
The University of Chicago

Lan Zhang
Carnegie Mellon University

The Review of Financial Studies Vol. 18, No. 2 © 2005

In theory, the sum of squares of log returns sampled at high frequency estimates their variance. When market microstructure noise is present but unaccounted for, however, we show that the optimal sampling frequency is finite and derives its closed-form expression. But even with optimal sampling, using say 5-min returns when transactions are recorded every second, a vast amount of data is discarded, in contradiction to basic statistical principles. We demonstrate that modeling the noise and using all the data is a better solution, even if one misspecifies the noise distribution. So the answer is: sample as often as possible.

A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data

Lan ZHANG, Per A. MYKLAND, and Yacine AÏT-SAHALIA

It is a common practice in finance to estimate volatility from the sum of frequently sampled squared returns. However, market microstructure poses challenges to this estimation approach, as evidenced by recent empirical studies in finance. The present work attempts to lay out theoretical grounds that reconcile continuous-time modeling and discrete-time samples. We propose an estimation approach that takes advantage of the rich sources in tick-by-tick data while preserving the continuous-time assumption on the underlying returns. Under our framework, it becomes clear why and where the "usual" volatility estimator fails when the returns are sampled at the highest frequencies. If the noise is asymptotically small, our work provides a way of finding the optimal sampling frequency. A better approach, the "two-scales estimator," works for any size of the noise.

Journal of the American Statistical Association
December 2005, Vol. 100, No. 472, Theory and Methods

Subsample

every kth value

$$m(k) \;=\; \text{floor}\left(\frac{N}{k}\right)$$

$$\begin{aligned}\mathrm{RV}(k) \;&=\; \sum_{j=1}^{m(k)} \left(y_{jk} - y_{(j-1)k}\right)^2 \\ &=\; \left(y_k - y_0\right)^2 \;+\; \left(y_{2k} - y_k\right)^2 \;+\; \cdots \;+\; \left(y_{mk} - y_{(m-1)k}\right)^2\end{aligned}$$

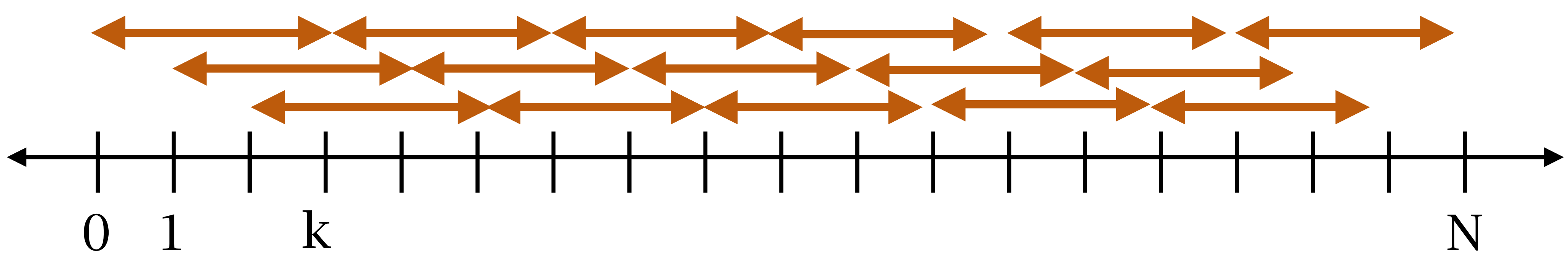
$$\mathrm{Var}(\xi) \;=\; \frac{\sigma^4 T^2 k}{N} \;+\; \frac{N\eta^4}{k}$$

Often choose k so $\quad \dfrac{\sigma^4 T^2 k}{N} \gg \dfrac{N\eta^4}{k} \qquad \Longleftrightarrow \qquad k \gg \dfrac{\eta^2}{\sigma^2 \tau}$

Microstructure noise small compared to intrinsic volatility

Subsample with overlapping intervals

$$
\begin{aligned}
RV(k) \;=\; & (y_k - y_0)^2 + (y_{k+1} - y_1)^2 \\
& + \cdots + (y_{N-1} - y_{N-k-1})^2 + (y_N - y_{N-k})^2
\end{aligned}
$$



Same mean, lower variance than non-overlapping, but hard to estimate variance (samples are not independent)

Multiscale (Ait-Sahalia, Mykland, Zhang)

Two different sampling frequencies

$$RV(N_1) = Q + 2N_1\eta^2 + \xi_1$$
$$RV(N_2) = Q + 2N_2\eta^2 + \xi_2$$

$$RV_* = \alpha_1 RV(N_1) + \alpha_2 RV(N_2)$$
$$\alpha_1 = \frac{N_2}{N_2 - N_1} \qquad \alpha_2 = -\frac{N_1}{N_2 - N_1}$$

Eliminate $\mathcal{O}(N)$ error term

Very dependent on precise structure of noise

High-Frequency Data and Volatility in Foreign-Exchange Rates

Journal of Business & Economic Statistics; Jan 1996; 14, 1; ABI/INFORM Global pg. 45

Bin ZHOU
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139

Exchange rates, like many other financial time series, display substantial heteroscedasticity. This poses obstacles in detecting trends and changes. Understanding volatility becomes extremely important in studying financial time series. Unfortunately, estimating volatility from low-frequency data, such as daily, weekly, or monthly observations, is very difficult. The recent availability of ultra-high-frequency observations, such as tick-by-tick data, to large financial institutions creates a new possibility for the analysis of volatile time series. This article uses tick-by-tick foreign-exchange rates to explore this new type of data. Unlike low-frequency data, high-frequency data have extremely high negative first-order autocorrelation in their return. In this article, I propose a model that can explain the negative autocorrelation and a volatility estimator for high-frequency data. The daily and hourly volatility estimates of exchange rate show some interesting patterns.

Because the mean of tick-by-tick return is negligible compared to the variance, I assume that the observation comes from process:

$$S(t) = B(\tau(t)) + \varepsilon_t. \tag{4}$$

Assuming that $\varepsilon_{t_i}, i = 1, \ldots, n$, are independent, I propose the following unbiased estimator of the volatility:

$$\hat{\sigma}_U^2 = \sum_{i=1}^{n} (X_i^2 + X_i X_{i-1} + X_{i+1} X_i), \tag{5}$$

where $X_i = S(t_i) - S(t_{i-1})$.

Zhou derivation

$$
\begin{aligned}
RV(y) &= \sum_{j=1}^{N} (y_j - y_{j-1})^2 \\
&= Q + 2N\eta^2 \;+\; \langle\text{noise}\rangle \\
ZV(y) &= \sum_{j=1}^{N-1} (y_j - y_{j-1})(y_{j+1} - y_j) \\
&= -N\eta^2 \;+\; \langle\text{noise}\rangle
\end{aligned}
$$

$$
\begin{aligned}
RV(y) \;+\; 2\,ZV(y) &= \sum_{j=1}^{N-1} \Big[(y_j - y_{j-1})^2 \\
&\qquad + 2(y_j - y_{j-1})(y_{j+1} - y_j) \Big] \\
&= Q \;+\; \langle\text{noise}\rangle
\end{aligned}
$$

(Zhou adds correction more symmetrically)

How to test variance estimators?

- Simulate stochastic process assumed in model

  Good agreement with theory (no surprise)
- Simulate artificial market

  Realistic microstructure

  Unrealistic long-term structure
- Apply statistical tests to real market data

  only real way to get meaningful test

Extensive empirical tests

Zero-intelligence realized variance estimation

Jim Gatheral · Roel C.A. Oomen

Finance Stoch (2010) 14: 249–283
DOI 10.1007/s00780-009-0120-1

Abstract Given a time series of intra-day tick-by-tick price data, how can realized variance be estimated? The obvious estimator—the sum of squared returns between trades—is biased by microstructure effects such as bid–ask bounce and so in the past, practitioners were advised to drop most of the data and sample at most every five minutes or so. Recently, however, numerous alternative estimators have been developed that make more efficient use of the available data and improve substantially over those based on sparsely sampled returns. Yet, from a practical viewpoint, the choice of which particular estimator to use is not a trivial one because the study of their relative merits has primarily focused on the speed of convergence to their asymptotic distributions, which in itself is not necessarily a reliable guide to finite sample performance (especially when the assumptions on the price or noise process are violated). In this paper we compare a comprehensive set of nineteen realized variance estimators using simulated data from an artificial "zero-intelligence" market that has been shown to mimic some key properties of actual markets. In evaluating the competing estimators, we concentrate on efficiency but also pay attention to implementation, practicality, and robustness. One of our key findings is that for scenarios frequently encountered in practice, the best variance estimator is not always the one suggested by theory. In fact, an ad hoc implementation of a subsampling estimator, realized kernel, or maximum likelihood realized variance, delivers the best overall result. We make firm practical recommendations on choosing and implementing a realized variance estimator, as well as data sampling.

Artificial market

2 The zero-intelligence limit-order book market

Motivated by our desire to model microstructure noise more realistically, we simulate from the limit-order book model of [25], SFGK hereafter. According to the specification of this model, limit orders can be placed at any integer price level p where $-\infty < p < \infty$. It may be natural to think of these price levels as being logarithms of the actual price (which must of course be non-negative). Limit sell orders may be placed at any level greater than the best bid $p^b(t)$ at time t and limit buy orders at any level less than the best offer $p^a(t)$. In particular, just as in real markets, limit orders may be placed inside the spread (if the current spread is greater than one tick). Market orders arrive randomly at rate μ, limit orders (per price level) arrive at rate α and a δ proportion of existing limit orders is canceled. All market orders and limit orders are for one share.

25. Smith, E., Farmer, J.D., Gillemot, L., Krishnamurthy, S.: Statistical theory of the continuous double auction. Quant. Finance **3**, 481–514 (2003)

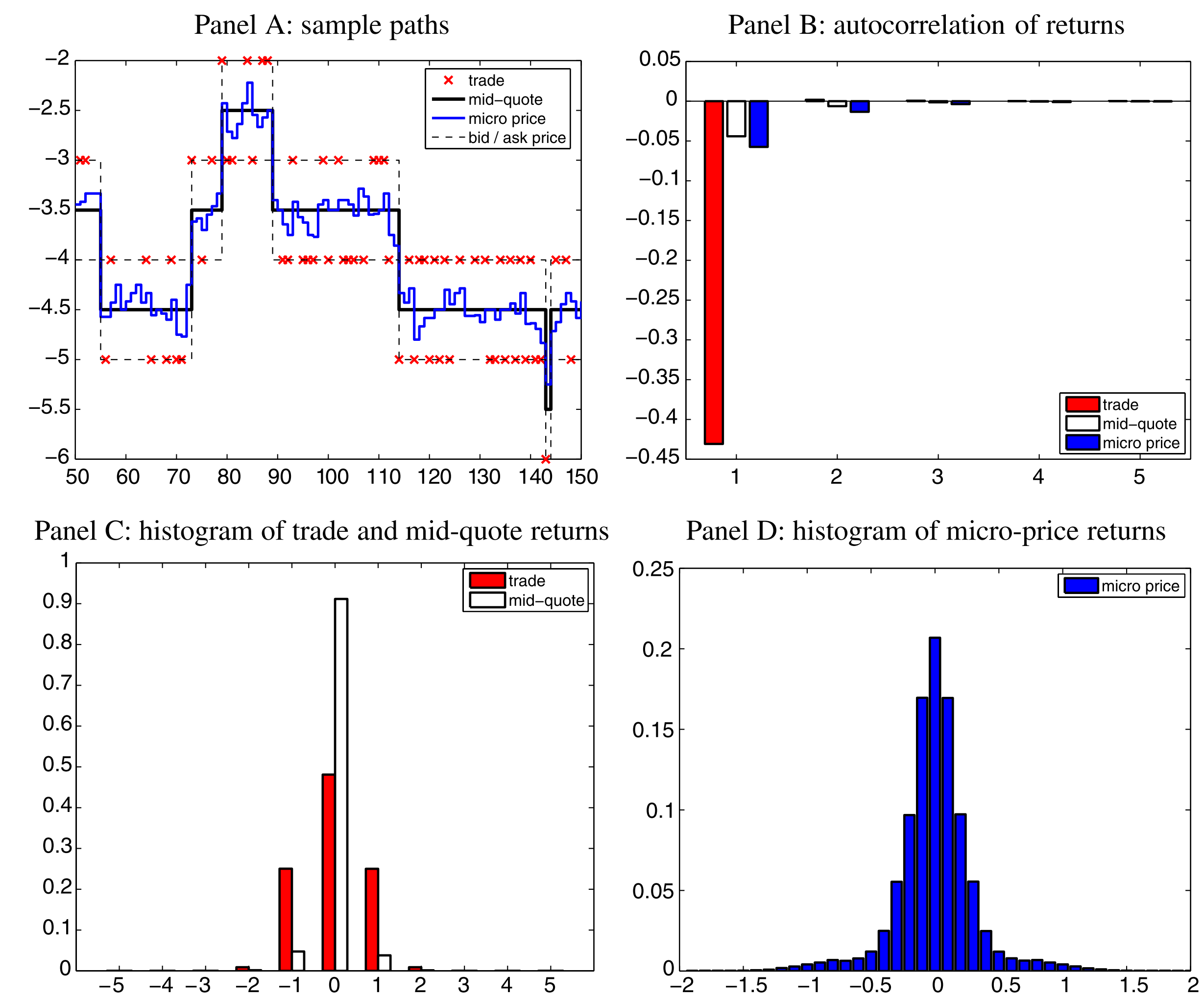


Fig. 1 Properties of the sampled price processes

This market has microstructure noise at short times At medium/long times, is exactly a random walk

Finance Stoch (2010) 14: 249–283
DOI 10.1007/s00780-009-0120-1

Gatheral/Oomen conclusions

To conclude, it should be stressed that the main virtue of the ad hoc selection of subsampling frequency or bandwidth, as suggested here, lies in its simplicity and robustness together with good all-round efficiency. Still, if one is willing to put in additional effort, there can be scope for refinement of this admittedly crude approach, albeit with no guarantee of success.

Finance Stoch (2010) 14: 249–283
DOI 10.1007/s00780-009-0120-1

Table 7 Ranking of alternative realized variance measures

	Trade prices			Quote prices			Micro-prices			loss		
$M =$	1,000	5,000	10,000	1,000	5,000	10,000	1,000	5,000	10,000	min	mean	max
1. Realized variance												
(a) highest ($q = 1$)	†	†	†	†	†	†	†	†	†	0.479	3.806	9.362
(b) ad hoc (5 mins)	†	†	†	†	†	†	†	†	†	0.674	1.853	2.513
(c) q^*_{RV}	†	†	†	✓	†	†	†	†	†	0.196	0.912	2.103
2. Bias-corrected RV of Zhou [28]												
(a) highest ($q = 1$)	†	†	†	✓	★	★	★	✓	†	0.009	0.535	1.422
(b) ad hoc (5 mins)	†	†	†	†	†	†	†	†	†	0.997	2.412	3.449
(c) q^*_{Zhou}	†	†	†	✓	★	★	★	✓	†	0.009	0.230	0.561
3. Two-scale RV of Zhang, Mykland, and Aït-Sahalia [27]												
(a) ad hoc ($q = 5$)	✓	✓	✓	✓	✓	✓	✓	★	★	0.007	0.097	0.179
(b) q^*_{ZMA}	†	✓	✓	✓	★	★	★	✓	†	0.006	0.152	0.382
4. Multi-scale RV of Zhang [26]												
(a) ad hoc ($q = 5$)	✓	✓	✓	✓	✓	✓	✓	★	★	0.006	0.098	0.191
(b) q^*_{Z}	✓	✓	✓	✓	★	★	★	✓	†	0.006	0.124	0.382
5. Realized kernel of Barndorff-Nielsen, Hansen, Lunde, and Shephard [10]												
(a) TH_2, ad hoc ($q = 5$)	✓	✓	✓	✓	★	★	✓	★	★	0.000	0.095	0.211
(b) TH_2, q^*_{BNLS}	✓	★	★	✓	★	★	★	✓	†	0.013	0.087	0.357
(c) TH_{16}, q^*_{BNLS}	★	★	★	✓	★	★	✓	✓	†	0.008	0.074	0.335
(d) Cubic, q^*_{BNLS}	✓	✓	✓	✓	★	★	★	✓	†	0.009	0.105	0.387
6. Fourier estimator of Malliavin and Mancino [20]												
(a) DIR kernel, q^*_F	†	†	†	✓	✓	✓	✓	†	†	0.108	0.375	0.683
(b) FEJ kernel, q^*_F	†	†	†	✓	✓	✓	✓	†	†	0.077	0.303	0.613
7. ALT of Large [19]	★	†	†	†	†	†	†	†	†	0.000	1.051	2.050
8. MLRV of AMZ [2]	★	★	★	★	★	★	★	✓	†	0.000	0.052	0.314
MLRV-MA(2)	✓	✓	✓	✓	✓	✓	✓	✓	★	0.033	0.082	0.151

Note A star (★) indicates the best performing RV measure, or those statistically indistinguishable at a 1% bootstrapped confidence level, a checkmark (✓) indicates an "acceptable" RV measure with an MSE at within 25% distance from the optimum, and a dagger (†) indicates a "bad" RV measure with an MSE at more than 25% distance from the optimum

No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications

Torben G. Andersen[a,c,*], Tim Bollerslev[b,c], Dobrislav Dobrev[a]

Journal of Econometrics 138 (2007) 125–180

We develop a sequential procedure to test the adequacy of jump-diffusion models for return distributions. We rely on intraday data and nonparametric volatility measures, along with a new jump detection technique and appropriate conditional moment tests, for assessing the import of jumps and leverage effects. A novel robust-to-jumps approach is utilized to alleviate microstructure frictions for realized volatility estimation. Size and power of the procedure are explored through Monte Carlo methods. Our empirical findings support the jump-diffusive representation for S&P500 futures returns but reveal it is critical to account for leverage effects and jumps to maintain the underlying semi-martingale assumption.

2.1.1. No leverage or jumps in the return generating process

even when the correct model is used to predict future return variability, in accordance with Eq. (6), the standardized returns will be fat-tailed relative to a Gaussian benchmark,

$$r(t) \cdot V^{-1/2}(t) \nsim \mathrm{N}(0,1). \tag{7}$$

On the other hand, returns normalized appropriately by the realized return variability are truly Gaussian,

$$r(t) \cdot \left[\int_0^t \sigma^2(s)\, \mathrm{d}s \right]^{-1/2} \sim \text{i.i.d. } \mathrm{N}(0,1). \tag{8}$$

With leverage, sample at times with equal variance.

2.1.2. The impact of leverage

There is compelling evidence that many markets, including those for equity indices, are characterized by a pronounced asymmetric relationship between return and volatility innovations. This is often labeled a "leverage effect" although this asymmetry arguably has little, if anything, to do with the underlying financial leverage; see, e.g., Campbell and Hentschel (1992), Bekaert and Wu (2000), and Bollerslev et al. (2005). In this case, the results from Section 2.1.1 are not valid. A key issue is whether this makes a practical difference for the distribution of standardized returns. A second question is whether it is feasible to restore distributional results for this scenario.

for a fixed positive period of "financial" time τ^*, we sample the logarithmic price process at times, $0 = t_0, t_1, \ldots, t_k, \ldots$, where the calendar time sampling points are defined by

$$t_k = \inf_{t>0}([r,r]_t - [r,r]_{t_{k-1}} > \tau^*), \quad k = 0, 1, \ldots, \tag{9}$$

so that returns are computed over intervals of identical quadratic variation, τ^*.

Denoting the sequence of returns sampled in financial time by

$$R_k \equiv p(t_k) - p(t_{k-1}), \quad k = 0, 1, 2, \ldots, \tag{10}$$

the following distributional result holds, even in the case of leverage,

$$R_k/\sqrt{\tau^*} \sim \text{i.i.d. N}(0,1), \quad k = 0, 1, 2, \ldots . \tag{11}$$

quantitativebrokers

INTRADAY REALIZED VARIANCE

LINWEI SHANG
APRIL 11, 2018

Volatility is a crutial building block in QB's execution system. We generate volatility profiles to guide our execution algorithms adjusting trading schedule. Volatility also plays a key role in transaction cost analysis since slippage is expected to be higher when market runs with high volatility.

The microstructuer noise in high-frequency data makes it a challenge to estimate volatility. Some standard estimators as realized variance (RV) are not reliable any more.

The problem becomes more difficult if we want to calculate RV for time periods as short as one minute and the estimator works for different financial instruments on different exchanges.

We assume that the latent price satisfies

$$\mathrm{d}X(t) = \sigma(t)\ \mathrm{d}w(t). \tag{1}$$

where $w(t)$ is a standard Brownian motion, $\sigma(t)$ is the instantaneous volatility function.

Then we can define the target we want to estimate, the integrated variance (IV),

$$\mathrm{IV} \equiv \int_a^b \sigma^2(t)\,\mathrm{d}t. \tag{2}$$

Any X(t) is
time-changed
Brownian motion

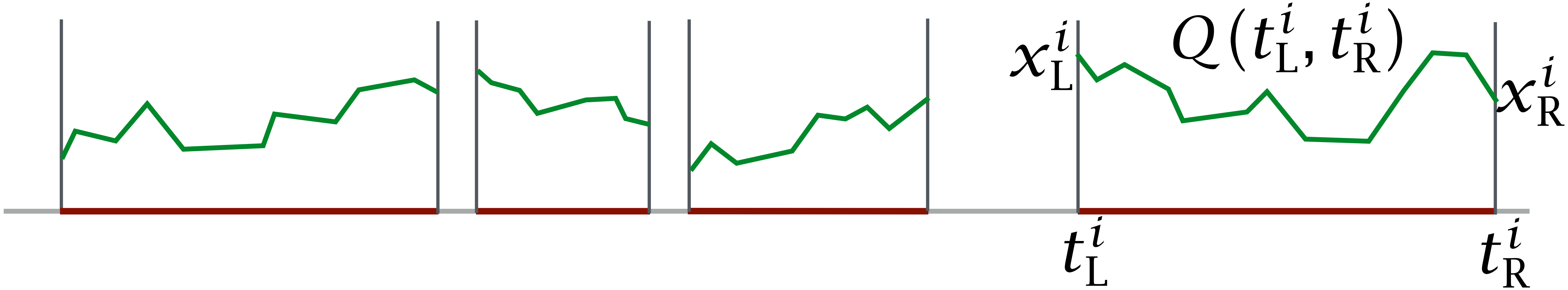
$$X(t) = B(\tau(t))$$

$X(t_{\mathrm{R}}) - X(t_{\mathrm{L}})$ is highly non-normal

$$\frac{X(t_{\mathrm{R}}) - X(t_{\mathrm{L}})}{\sqrt{\tau(t_{\mathrm{R}}) - \tau(t_{\mathrm{L}})}}$$ is unit normal
(ignoring leverage)

Idea:

- Measure $Q(t_\mathrm{L}, t_\mathrm{R})$ on many historical intervals using a candidate estimator

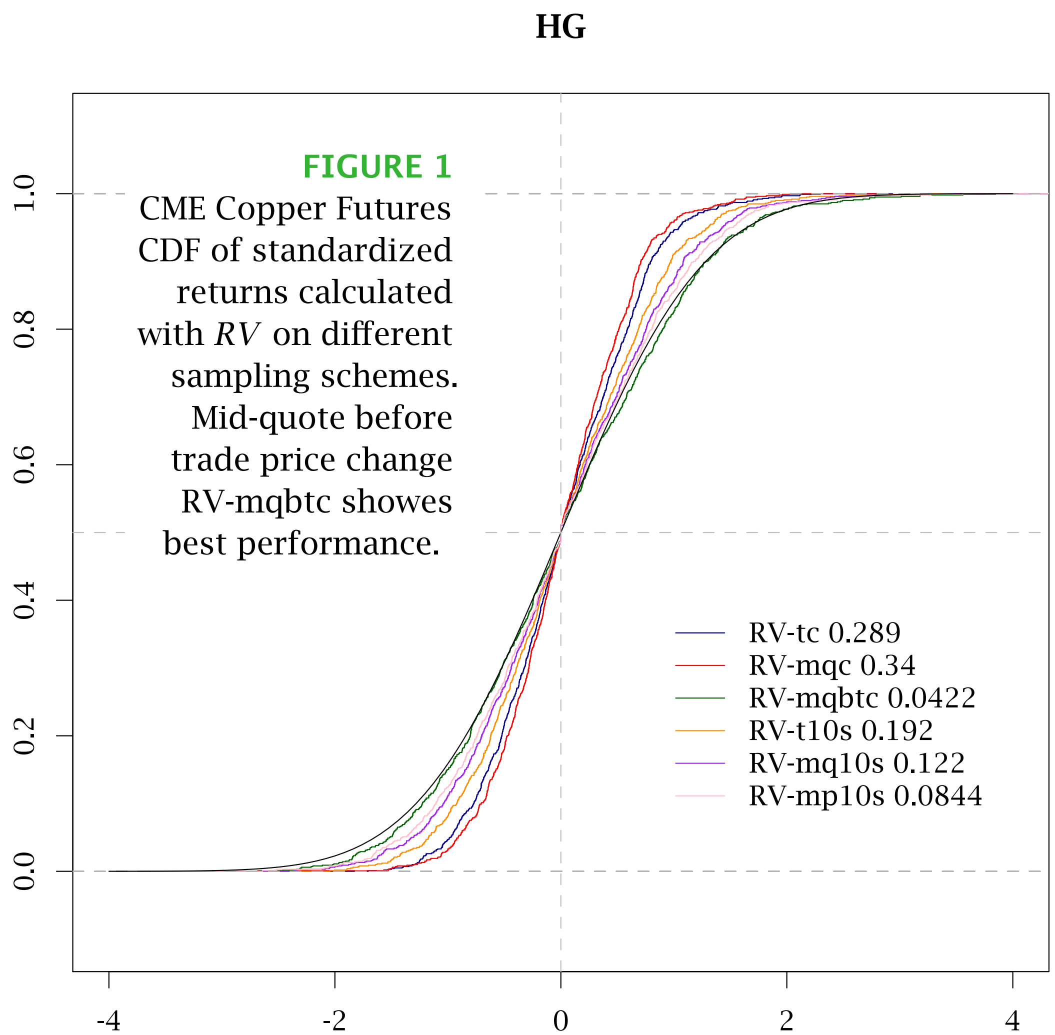


See if scaled price changes reduce to a unit normal population

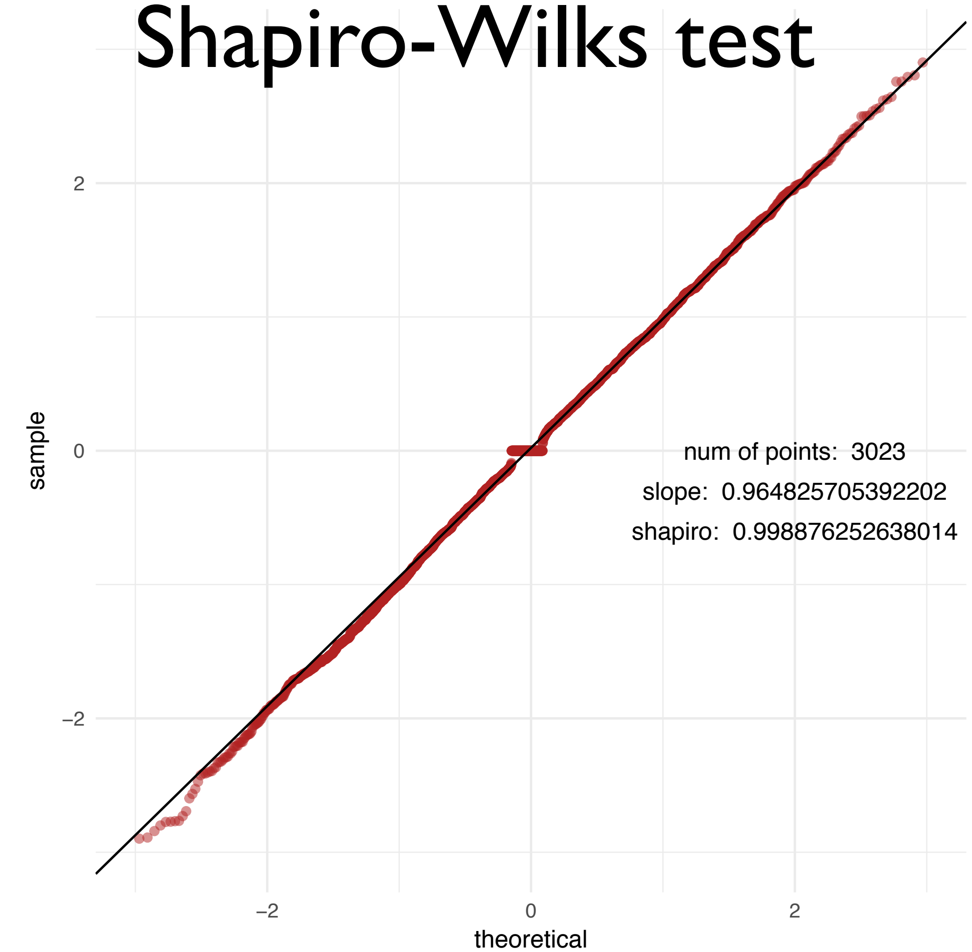
$$z_i = \frac{x_\mathrm{R}^i - x_\mathrm{L}^i}{\sqrt{Q(t_\mathrm{L}^i, t_\mathrm{R}^i)}}$$

This avoids the problem of forecasting or modeling Q

Normality test



Shapiro-Wilks test



- `tc` : Trade prices when it change
- `mqc` : Mid-quotes when it change
- `mqbt` : Mid-quotes when trade prices change
- `t10s` : Trade prices every 10 seconds
- `mq10s` : Mid-quotes every 10 seconds
- `mp10s` : Micro-price every 10 seconds

BIAS CORRECTION

Since any sampling scheme and any bias correction device may compose a valid estimator, the estimator candidates can form a big matrix with dozens of entries. It will be hard to distinguish one best from them. Therefore, we will first try all the bias correction methods on `mqbtc` which shown to be the best in previous part.

As can be seen from Table 2, RV-mqbtc gives minimum Wasserstein metric on 29 of 84 products which is again the highest. But in the second part of the test, TSRV-mqbtc gives lowest mean score. And all bias correction methods except for KTH_{16} can improve performance on average. But the improvement is not significant and no specific methods can consistently precede the original RV estimator.

Estimator	Freq
RV-mqbtc	29
KCubic-mqbtc	13
MSRV-mqbtc	12
KTH_2-mqbtc	9
ZHOU-mqbtc	9
KTH_{16}-mqbtc	7
TSRV-mqbtc	5

Estimator	Score
TSRV-mqbtc	1.2847
MSRV-mqbtc	1.3097
KTH_2-mqbtc	1.3562
KCubic-mqbtc	1.3706
ZHOU-mqbtc	1.3791
RV-mqbtc	1.4201
KTH_{16}-mqbtc	7.5281

Conclusion from realized variance study

TABLE 3
Results of all estimators.

Estimator	Freq
RV-mqbtc	20
RV-mp10s	16
TSRV-mqc	12
TSRV-tc	6
MSRV-mqbtc	5
KTH_2-mqbtc	5
TSRV-mp10s	4
ZHOU-mqbtc	3
KCubic-mqbtc	3
TSRV-mq10s	3
TSRV-mqbtc	3
RV-mqc	1
RV-t10s	1
RV-mq10s	1
TSRV-t10s	1
RV-tc	0
KTH_{16}-mqbtc	0

Estimator	Score
TSRV-mqbtc	1.4692
MSRV-mqbtc	1.5000
TSRV-mqc	1.5434
KTH_2-mqbtc	1.5458
KCubic-mqbtc	1.5556
ZHOU-mqbtc	1.5896
RV-mqbtc	1.6688
TSRV-mp10s	1.7561
TSRV-tc	1.8253
TSRV-mq10s	1.8360
RV-mp10s	2.0808
RV-mq10s	2.7010
TSRV-t10s	2.8933
RV-t10s	4.2105
RV-tc	4.2519
RV-mqc	5.7916
KTH_{16}-mqbtc	9.3489

`RV-mqbtc` is much more easier to be implemented than `TSRV-mqbtc`, Therefore, RV-mqbtc is our final selection.

Recommendation

- Zhou is simple and robust
- Garman-Klass works well if can get max/min
- More complex correction schemes do not work well
- Simplest and best:
  quote midpoint at times of trades with new prices
  (Linwei Shang, QB)

Correlation

- Same issues as volatility
- Less practical use (except for risk management)
- Must sample on uniform time grid
  to have consistent set of time points
- Correlation decreases at short lags (Epps effect)

Theory is available

High-Frequency Covariance Estimates With Noisy and Asynchronous Financial Data

Yacine AÏT-SAHALIA, Jianqing FAN, and Dacheng XIU

Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Theory and Methods
DOI: 10.1198/jasa.2010.tm10163

This article proposes a consistent and efficient estimator of the high-frequency covariance (quadratic covariation) of two arbitrary assets, observed asynchronously with market microstructure noise. This estimator is built on the marriage of the quasi–maximum likelihood estimator of the quadratic variation and the proposed generalized synchronization scheme and thus is not influenced by the Epps effect. Moreover, the estimation procedure is free of tuning parameters or bandwidths and is readily implementable. Monte Carlo simulations show the advantage of this estimator by comparing it with a variety of estimators with specific synchronization methods. The empirical studies of six foreign exchange future contracts illustrate the time-varying correlations of the currencies during the 2008 global financial crisis, demonstrating the similarities and differences in their roles as key currencies in the global market.

KEY WORDS: Covariance; Generalized synchronization method; Market microstructure noise; Quasi-Maximum Likelihood Estimator; Refresh Time.

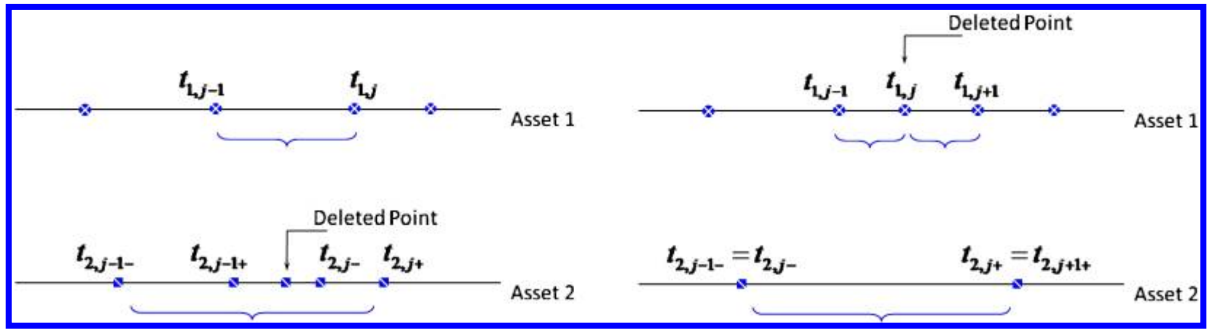


Figure 1. These two graphs illustrate how the HY synchronization method delete data points. The online version of this figure is in color.

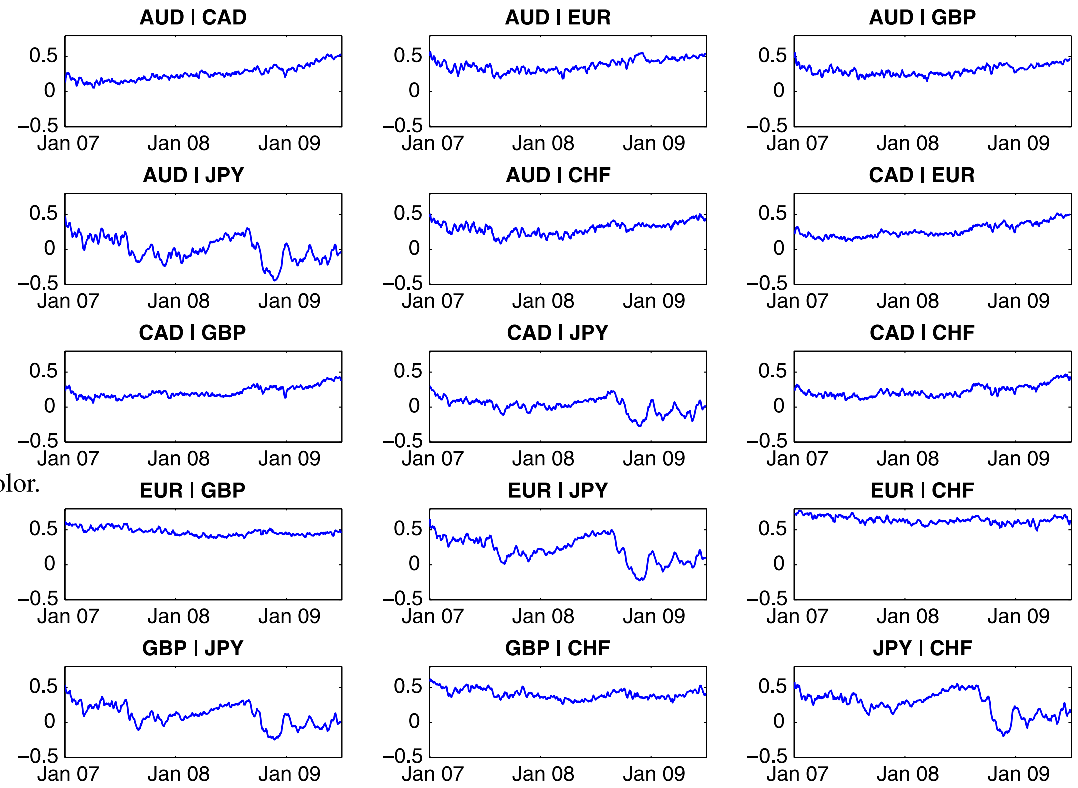


Figure 4. In this figure, we report the time series of correlation estimates of all six FX futures. Each curve is a 5-day moving average o
daily correlations. The online version of this figure is in color.

Epps effect

Comovements in Stock Prices in the Very Short Run

THOMAS W. EPPS*

Correlations among price changes in common stocks of companies in one industry are found to decrease with the length of the interval for which the price changes are measured. This phenomenon seems to be caused by nonstationarity of security price changes and by the existence of correlations between price changes in the same stock—and in different stocks—in successive periods. Although such correlations are not necessarily inconsistent with market efficiency, the data do reveal the presence of lags of an hour or more in the adjustment of stock prices to information relevant to the industry.

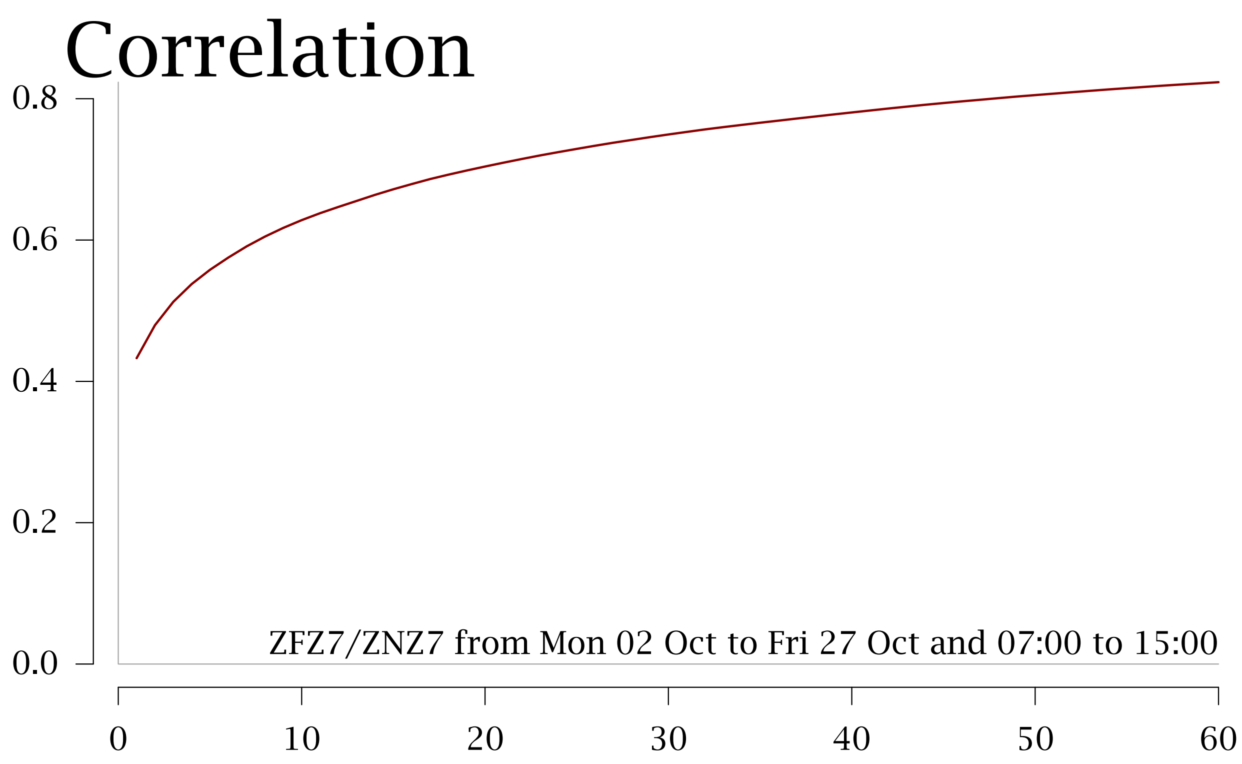
KEY WORDS: Efficient-markets hypothesis; Portfolio theory; Financial markets; Fisher effect.

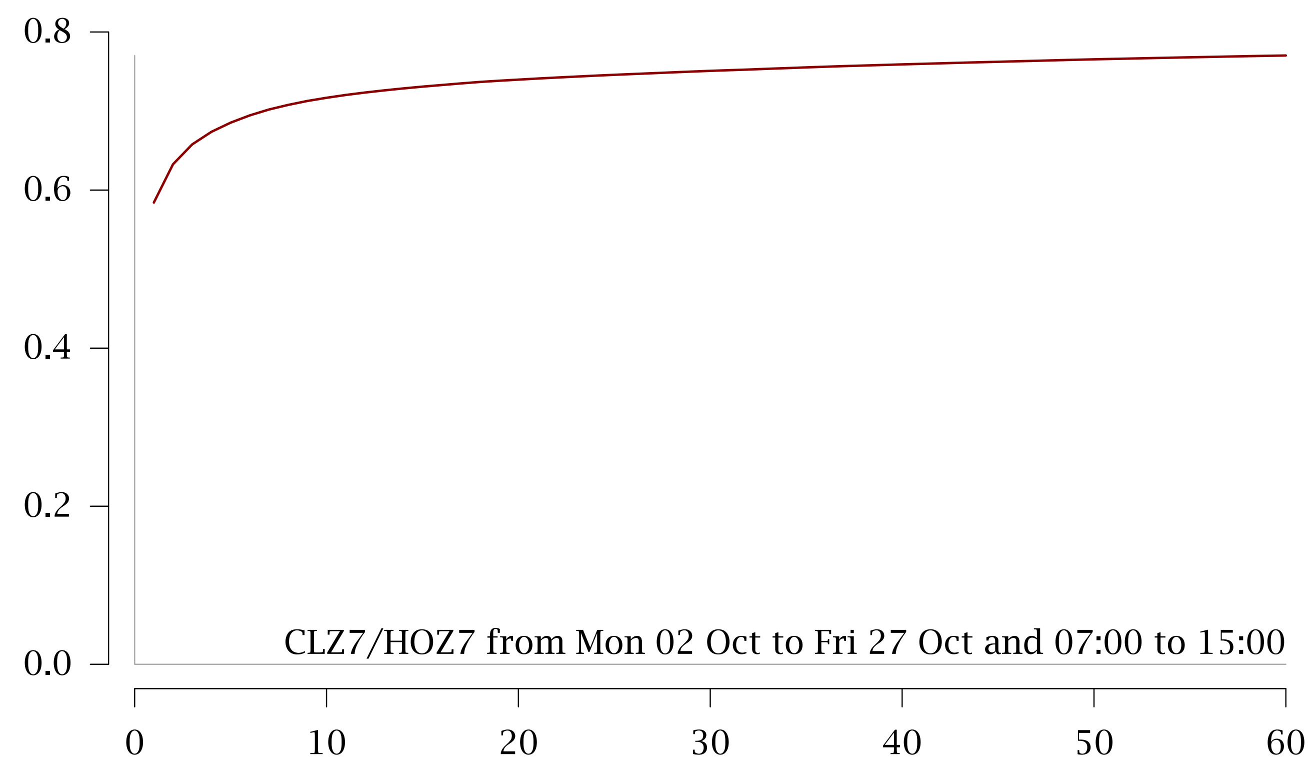
© Journal of the American Statistical Association
June 1979, Volume 74, Number 366
Applications Section

1. Correlations of Changes in Log Price for Four Stocks During Intervals of 10 Minutes to Three Days

	Pairs of Stocks					
Interval	*AMC-Chrysler*	*AMC-Ford*	*AMC-GM*	*Chrysler-Ford*	*Chrysler-GM*	*Ford-GM*
10 minutes	001	.009	− 009	− 014	007	055
20 minutes	009	018	011	017	026	118
40 minutes	.006	012	.014	041	040	197
One hour	−.043	.057	064	.023	065	.294
Two hours	029	.060	094	.112	129	383
Three hours	031	.158	111	361	518	519
One day	−.067	.170	078	342	442	571
Two days	− 020	223	186	336	449	572
Three days	− 098	203	.100	334	542	645

Epps effect in futures (HW3)





Time lag in seconds