Returns, Durations, and Time Endogeneity

Denis Pelletier   Qifeng Weng

NC State University   NC State University

Job Market Talk @ NC State University

Dec. 16, 2015

Outline

What is Time Endogeneity Effect?

Definition: Time endogeneity is a situation where the instantaneous correlation between the irregular spaced time intervals ($d_i \equiv t_i - t_{i-1}$) and Wiener terms (W_{t_i}) in a log-price process is nonzero.

Intuition: think about the leverage effect.

Prevailing Realized Volatility (RV) estimators are built under the assumption that observation time being exogenous. Hence, time endogeneity violates the assumption and generates the bias.

Biased RV Problem

A standard approach modeling the asset price in the continuous time literature is using an Itô process:

$$d \ln S_t = \mu_t dt + \sigma_t dW_t.$$

- W_t is a Wiener process. Its increment follows a Normal distribution.
- This log-price process can be extended to encompass: **price jumps** and **market microstructure noise**.
- The similar setting is used to model option prices, interest rate, exchange rate, *etc.*

Return

Thus the increment of $\ln S_t$ at the observation time t_i is defined as

$$\Delta X_i \equiv \ln S_{t_i} - \ln S_{t_{i-1}}.$$

By design, we see that the difference of log-prices, ΔX_i, is also the continuously compounded rate of return of the i-th trade (log-return).

Underlying Time Grid

transaction: $(i-1)^{th}$ trade $\qquad\qquad$ i^{th} trade

time: $\underbrace{t_{i-1} \qquad\qquad\qquad t_i}$

duration: $d_i \equiv t_i - t_{i-1}$

Price level: $S_{t_{i-1}} \qquad\qquad\qquad S_{t_i}$

Return: $\Delta X_i \equiv \ln S_{t_i} - \ln S_{t_{i-1}}$

Integrated Volatility (IV)

A natural way of defining the cumulative variance of asset prices over successive time periods in this model family is simply to integrate the instantaneous latent volatility over the time. It is called **Integrated Volatility (IV)** in literature, defined as:

$$\mathrm{IV} \equiv \int_0^T \sigma_t^2 dt.$$

Realized Volatility (RV)

Realized Volatility (RV) is simply the sum of squared returns, defined as:

$$[X, X]_t = \sum_{t_i \leq t} (\Delta X_i)^2$$

where ΔX_i is log-returns.

- Jacod and Protter (1998) proves (when time is exogenous):

$$\mathrm{plim}_{n \to \infty} \mathrm{RV} = \mathrm{IV}.$$

- **Empirical Value**: Accurate equity volatility proxy (Realized Volatility) serves as an essential role in managing portfolios, reducing option pricing error (Christoffersen *et al.*, 2014), and improving VaR forecasts (Louzis *et al.*, 2014).

Li, Mykland *et al.* (2013) Critique

1. RV overestimates IV when time is endogenous:

$$N^{1/2}(\mathrm{RV}-\mathrm{IV}) \to \underbrace{\frac{2}{3}\int_0^t v_s\sigma_s dX_s}_{\text{asymptotic bias}} + \int_0^t \sqrt{\frac{2}{3}u_s - (\frac{2}{3}v_s)^2}\sigma_s^2 dB_s.$$

   Note: It is very complicated to estimate this bias term.

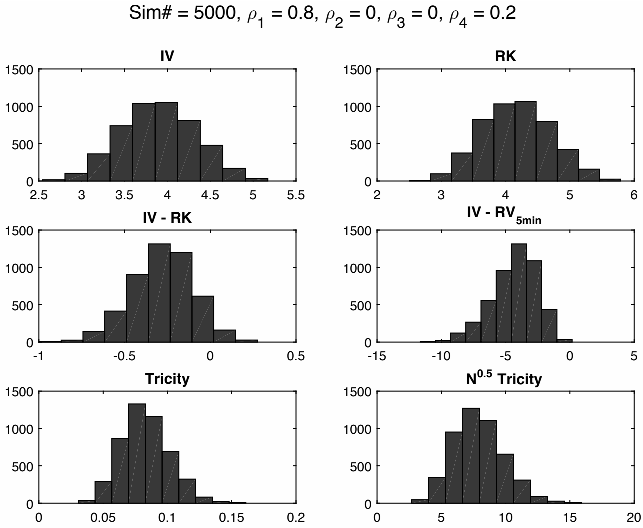
2. The tight connection between the biased term and duration:

$$N^{1/2}\sum_{t_i\le t}\Delta d_i \Delta X_i \xrightarrow{p} \frac{1}{3}\int_0^t v_s\sigma_s ds.$$

3. An expressive symptom of the time endogeneity effect, nonzero **Tricity**: $[X,X,X]_t \equiv \sum_{t_i\le t}(\Delta X_i)^3 \nrightarrow 0$.

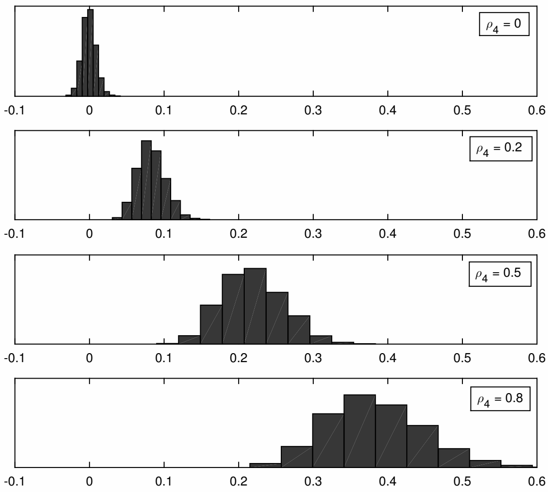
4. A CLT for tricity has been developed to prove the existence of time endogeneity effect in the real data.

Simulation Evidence I



Note: ρ_1 controls the leverage effect. ρ_4 brings in the time endogeneity effect.

Simulation Evidence II: Tricity Plots

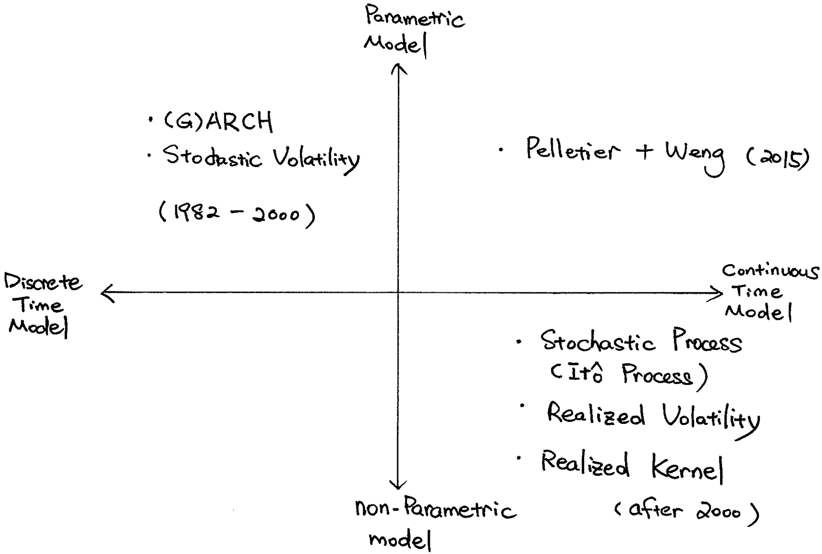


Note: As time endogeneity effect grows, tricity deviates further from zero and has more dispersion

Objective of Our Paper

To build a parametric model that produces an asymptotically **unbiased** volatility proxy for equity price given time is endogenous. Also it has in-model volatility **forecasting** ability.

A Brief History of the Volatility Literature



Parametric Discrete Time Model Literature

Discrete Time Volatility Models:

- ARCH model (Engle, 1982); GARCH model (Bollerslev, 1986).
- Competitive alternative: Stochastic Volatility (SV) model (Hull & White, 1987; Kim *et al.*, 1998).

Limitations of these types of model:

- Due to the model assumptions, they can only accommodate equidistant frequency data (normally daily data).
- Aggregation of high-frequency data leads to the loss of temporal information, *etc.*
- Other challenging issues like the difficulty of capturing long-memory dependenies and handling multivariate system (curse of dimensionality).

Nonparametric Continuous Time Model Literature

- Jacod and Protter (1998) proves $\text{plim}_{n\to\infty} RV = \text{IV}$
- Barndorff-Nielsen and Shephard (2001, 2002) introduced this theory to financial econometrics.
- More than 400 realized volatility estimators tackling price jumps and microstructure noise issues have been developed since then.

Prevailing Realized Volatility Estimators

1. **RV$_{\mathbf{5min}}$** (Barndorff-Nielsen and Shephard, 2001); It aggregates all returns within a 5 minute calendar time interval as 5-min return, denoted as $\Delta X_i^{5\text{min}}$ to mitigate market microstructure noise.

2. **TRV** (Mancini, 2009) is also designed to eliminate jumps via sophisticatedly selected thresholds.

3. **BPV** (Barndorff-Nielsen and Shephard, 2004) is designed to remove price jump effects by summing over the product of absolute returns from two neighboring periods.

4. **Realized Kernel** (RK) (Barndorff-Nielsen *et al.* 2011) serves as a robust volatility estimator to both price jumps and microstructure noise.

Model Overview

Our model:

- has two inputs: high frequency equity trading prices (eventually transformed to returns) and their corresponding durations between consecutive transactions.
- has one output: an unbiased volatility estimator.
- has four layers.
- encompasses: time exogeneity effect, time endogeneity effect and leverage effect.

Asset Price Processes

The first layer is a Wiener process of logarithmic price.

The dynamics of the logarithmic price S_t at time t is defined as:

$$d \ln S_t = \sigma_t dW_{0,t},$$

where $W_{0,t}$ is a Weiner term. σ_t is a latent instantaneous volatility as its coefficient. We ignore the drift term in our model for simplicity. For future work we will include price jumps and market microstructure noise.

Intensity

The second layer is a Poisson process for durations.

λ_{t_i} is the stochastic **intensity** characterizing a point process. Intuitively, λ_{t_i} can be interpreted as the instantaneous probability to have a i-th trade occur at time t. It is defined as:

$$\lambda_{t_i} = \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} Prob[N(t + \Delta t) - N(t) = 1 | F_{t-}] \right),$$

where $N(t) = \sum_{i \geq 1} 1(T_i \leq t)$ is a counting process, summing up the total numbers of events up to and including time t. F_{t-} is all information available up to time t.

Duration

Naturally, we define the time span between the $(i-1)$-th trade and the i-th trade as **duration** d_i:

$$d_i \equiv T_i - T_{i-1},$$

where T_i is the instant of the i-th occurrence of an event, satisfying $0 < T_1 < T_2 < \cdots$.

Duration Poisson Process

We assume d_i follows a conditional exponential distribution with mean $\lambda_{t_{i-1}}^{-1}$:

$$f(d_i|\lambda_{t_{i-1}}) = \frac{1}{\lambda_{t_{i-1}}} \cdot \nu_i \qquad \text{with} \quad \nu_i \sim \text{Exp}(1).$$

We could also write it as $f(d_i|\lambda_{t_{i-1}}) \sim \text{Exp}(\lambda_{t_{i-1}})$, which is the definition of a Poisson process.

Intensity-Volatility OU Process

The third layer in our model is a bivariate Ornstein-Uhlenbeck (OU) process for two latent variables: log-intensity λ_t and log-variance σ_t^2, which is denoted as:

$$dX_t = A(\mu - X_t)dt + SdW_{-0,t}$$

where:

$$X_{t_i} = \begin{bmatrix} \ln \lambda_t \\ \ln \sigma_t^2 \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$
$$S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad W_{-0,t} = \begin{bmatrix} W_{1,t} \\ W_{2,t} \end{bmatrix}.$$

OU process:(i) It has a closed form solution. (ii) It can be interpreted as the continuous time version of the AR(1) process. (iii) It is a mean-reverting process given A is PD.

Time Endogeneity Setting

The last layer of our model introduces the endogenous time effect to our model and links aforementioned ingredients all together via a Gaussian Copulas.

The idea of Copulas is that joint distribution function can be written as a unique Copulas function of marginal distributions of each random variables under the condition that all marginal distributions are continuous (Sklar, (1959)).

We impose contemporaneous correlations, denoted as ρ, to our model.

- ρ_1: the instantaneous correlation between Wiener terms in $\ln S_{t_i}$ and $\ln \lambda_{t_i}$ dynamics. It is defined as:

$$\rho_1 \equiv < W_{0,t}, W_{1,t} > .$$

- **Leverage Effect** parameter ρ_2: the instantaneous correlation between Wiener terms in $\ln S_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics. It is defined as:

$$\rho_2 \equiv < W_{0,t}, W_{2,t} > .$$

- ρ_3 is the instantaneous correlation between Wiener terms in $\ln \lambda_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics. It is defined as:

$$\rho_3 \equiv < W_{1,t}, W_{2,t} > .$$

- **Endogenous Time Effect** parameter ρ_4:

$$\rho_4 \equiv \mathrm{Corr}(\Phi^{-1}(F_{\exp}(\nu_i)), \Delta W_{0,t_i}/\sqrt{d_i}).$$

Gaussian Copulas

The Gaussian copulas **density** function in our case is defined as:

$$c(x) = \frac{1}{|\Gamma|^{1/2}} \exp\left[-\frac{1}{2}u'(\Gamma^{-1} - I_4)u\right],$$

where:

$$x = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \nu_i \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_4 \\ \rho_1 & 1 & \rho_3 & 0 \\ \rho_2 & \rho_3 & 1 & 0 \\ \rho_4 & 0 & 0 & 1 \end{bmatrix}, \quad u = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \Phi^{-1}(F_{\exp}\nu_i) \end{bmatrix}.$$

The normal quantile function u_i is given by: $u_i = \Phi^{-1}(F_i(x_i))$, where $F_i(\cdot)$ is the corresponding cumulative density function (cdf) of x_i.

Overview of Estimation Procedure

1. Discretize and linearize these continuous time processes.
2. Transformed the model to the state space representation.
3. Conduct Quasi-Maximum Likelihood Estimation (QMLE) via the Kalman filter.

Discretization and Linearization

1. **Log-price Process**:
   Discretization:

   $$y_{t_{i+1}} \equiv \ln S_{t_{i+1}} - \ln S_{t_i} = \sigma_{t_i} \cdot \zeta_{t_{i+1}} \quad \text{with} \quad \zeta_{t_{i+1}} \sim N(0, d_{i+1}).$$

   Linearization:

   $$\ln(y^2_{t_{i+1}}) = \ln(\sigma^2_{t_i}) + \ln(d_{i+1}) + \ln(\eta^2_{t_{i+1}}).$$

   We employ Euler Scheme here, which introduces discretization errors.
2. **Duration Poisson Process** is linearized as:

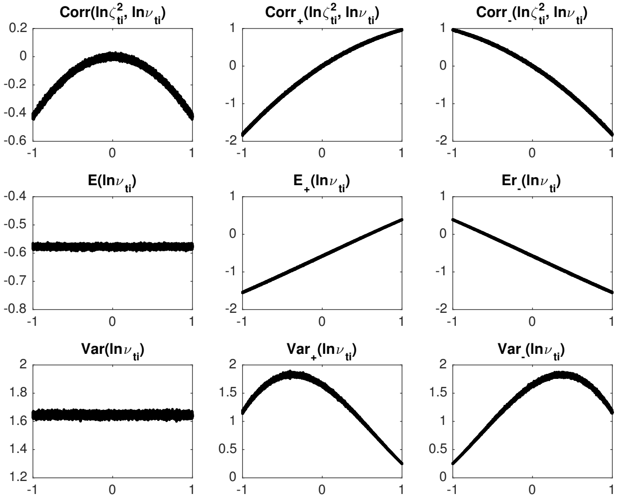
   $$\ln(d_i) = -\ln(\lambda_{t_{i-1}}) + \ln \nu_i.$$

3. **Intensity-Volatility OU Process** has a closed form solution:

   $$X_{t_{i+1}} = [I_2 - \mathrm{expm}(-Ad_{i+1})]\mu + \mathrm{expm}(-Ad_{i+1})X_{t_i} + \Omega_{t_{i+1}},$$

   Using this solution won't bring in any errors. It is an AR(1).

Identification Issues of ρ_4 When Imposing QMLE



The Kalman Filter

Finally, we got the state space representation of our model conditioning on the sign of the return:

Observation Equation:

$$Y_{t_i} = \begin{bmatrix} \mu^{\psi}_{*,t_i} \\ -1.2704 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X_{t_{i-1}} + \begin{bmatrix} \psi_{*,t_i} \\ \varepsilon_{*,t_i} \end{bmatrix},$$

State Equation:

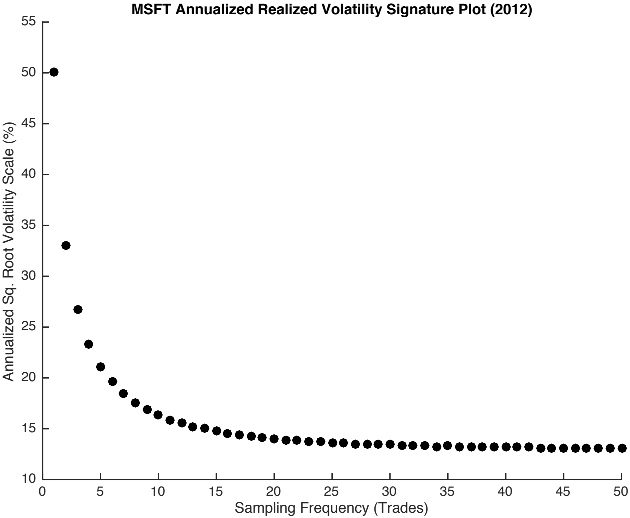
$$X_{t_i} = [I_2 - \text{expm}(-Ad_i)]\mu + \mu^{\Omega}_{*,t_i} + \text{expm}(-Ad_i)X_{t_{i-1}} + \Omega_{*,t_i},$$

Data Description

Table: MSFT Tick Data Summary Statistics (1/16/2012 - 2/15/2012)

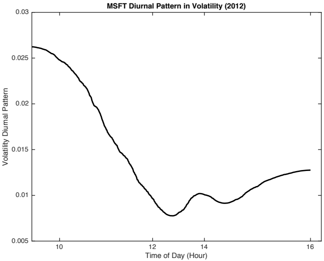
	Return 1	Return 50	Duration 1	Duration 50
Obs.	4,393,504	87,753	4,393,504	87,753
Mean	0.000000	0.000002	0.436730	21.862079
S.E.	0.00014	0.00032	1.03410	17.69725
Min	-0.01265	-0.00445	0.00300	0.16600
Max	0.01269	0.00332	30.88000	159.08000
Median	0.00000	0.00000	0.03000	17.40400
Skewness	0.01893	-0.02774	4.94251	1.47790
Kurtosis	94.446	7.741	41.910	6.014
ACF(1)	-0.37329	-0.04921	0.23906	0.48345
ACF(2)	-0.05921	0.01390	0.19626	0.38865
ACF(3)	-0.01881	0.00182	0.16909	0.36563
ACF(4)	-0.00624	0.00578	0.15169	0.35625
ACF(5)	-0.00300	-0.01215	0.13832	0.35701

Annualized RV Signature Plot by Tick-time Sampling Scheme

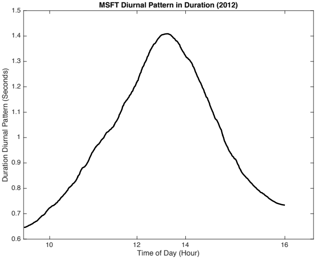


Diurnal Pattern

Traders have a tendency to trade more frequently near the opening of the market due to the overnight effect as well as near the closing time, and less frequently in the middle of the day. We need to remove the diurnal pattern from data to avoid model misspecification



(a) Volatility Diurnal Pattern



(b) Duration Diurnal Pattern

Estimation Results

Table: Estimation Results for MSFT Tick Data

	Individual Trade		15-Trade		50-Trade		100-Trade	
Parameters	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
A_{11}	0.53684	0.00001	0.00436	0.07959	0.04870	0.00945	0.02334	0.01364
A_{12}	-0.10675	0.00000	0.00594	0.01613	0.00146	0.00008	0.00057	0.00022
A_{21}	-0.99974	0.00001	0.07056	0.16153	0.08886	0.02794	0.02632	0.00033
A_{22}	18.23772	0.00002	0.33592	0.43609	0.04188	0.00199	0.00774	0.00000
σ_d	0.71824	0.00001	0.08642	0.06885	0.03791	0.01206	0.02983	0.00502
σ_v	56.51033	0.00001	4.79010	0.21194	1.35651	0.02616	0.49807	0.00925
μ_d	0.82357	0.00000	-2.34626	0.00788	-3.55998	0.03922	-4.28808	0.01232
μ_v	-15.01144	0.00002	-12.49792	0.33589	-12.09036	0.02097	-11.88744	0.00971
ρ_1	0.05720	0.00002	0.05296	0.38421	0.11060	0.02398	0.07236	0.01517
ρ_2	0.01046	0.00001	0.01360	0.20994	-0.04573	0.00545	-0.12913	0.01153
ρ_3	-0.43701	0.00001	0.97416	0.79751	0.98770	0.00814	0.97939	0.01232
ρ_4	-0.03252	0.00000	0.01513	0.16580	0.07309	0.01551	0.11891	0.02785

Forecasting Accuracy Test

By using loss function family proposed by Patton (2013), and implementing DMW test, we have:

Table: Comparison of proposed model and AR(1) model forecasts

Loss Function	Realized Kernel	Daily Squared Return	5-min RV
$b = 1$	1.12	2.84	-12.79
$b = 0$ (MSE)	1.23	2.82	-12.18
$b = -1$	1.27	2.76	-11.53
$b = -2$ (QLIKE)	1.11	2.60	-9.85
$b = -5$	-0.57	0.56	-2.01

Note: (a) both MSE loss and QLIKE loss reach the same testing results. (b) An absolute value of test score > 1.96 indicates the statistically significance. (c) Positive sign implies that test results are in favor of our model.

Conclusion

1. Our model produces a unbiased volatility estimator given time is endogenous.

2. Our model demonstrates impressive prediction power through forecast accuracy tests.