A Stochastic Price Duration Model for Estimating High-Frequency Volatility

Wei Wei[*]

Denis Pelletier[†]

Abstract

We propose a class of stochastic price duration models to estimate high-frequency volatility. A price duration measures how long it takes for the asset price to change by a given amount. It is directly linked to volatility from the passage time theory for Brownian motions. Modeling with price durations renders more effecient sampling scheme compared to return-based estimators. Also, our parametric approach allows us to estimate intraday spot volatility and incorporate additional information such as trade durations.

1 Introduction

Volatility plays a pivotal role in modern day financial economics. Since asset prices are generally considered to be driven by Brownian motions, the natural way to estimate volatility is to look at how much price changes in a given time interval. In particular,

[*] Department of Economics, North Carolina State University, Raleigh, NC 27695, USA, email: wwei6@ncsu.edu

[†] Department of Economics, North Carolina State University, Raleigh, NC 27695, USA, email: denis_pelletier@ncsu.edu

if volatility is constant, the variance of the return is a linear function of volatility. The passage time theory for Brownian Motion provides another approach to estimate volatility: one can look at how long it takes for the price to change by a given amount. Let price duration refer to the waiting time for the logarithmic price to travel the distance δ. The expectation of the price duration is related to the inverse of volatility. Intuitively, if volatility is high, price will be changing quickly and the expected price duration will be relatively short. While most volatility estimation methods are based on returns, this paper utilizes price durations to model high-frequency volatility.

The time-varying nature of volatility poses challenges to its estimation. Roughly speaking, there are three return-based methods to estimate volatility, namely GARCH, stochastic volatility(SV) and realized volatility (RV). GARCH-type models[1] assume that volatility is some function of past returns. In the SV-type models, volatility is assumed to be random and to follow a stochastic process. The availability of high-frequency financial data has popularized the RV estimator, which uses returns sampled at shorter horizons (such as 5 minutes) to measure volatility at a longer horizon (such as a day). The RV approach assumes volatility to be stochastic without specifying any parametric form. In the frictionless arbitrage-free world, the sum of squared returns converges in probability to integrated volatility when the sampling frequency goes to infinity. However, since observed prices are contaminated by market microstructure noise, realized volatility is a biased estimator of the actual volatility, and the problem becomes more severe when sampling frequency increases. One solution is to sample sparsely[2]. The optimal sampling frequency can be determined by considering the trade-off between the bias induced by microstructure noise and the variance induced by decreasing the sampling frequency. In

[1] See Hansen and Lunde (2005) for a list of 330 specifications in the GARCH universe and their evaluation.

[2] There are more sophisticated ways to deal with market microstructure noise, such as subsampling (see Zhang, Mykland, and Ait-Sahalia, 2005), pre-averaging (see Jacod, Li, Mykland, Podolskij, and Vetter, 2009) or realized kernels (see **?**).

practice, 5-minute RV is commonly used.

The literature on duration-based volatility estimation is considerably smaller and most of the work employs the autoregressive conditional duration (ACD) model. Engle and Russell (1998) propose ACD to model the durations between trades. They also apply the model to price durations by treating the price arrival times as a point process, and link the price arriving intensity to volatility. ACD is similar to GARCH: the volatility traced out from price intensity is assumed to be deterministic. Tse and Yang (2012) adopt the augmented ACD specification to model price durations and estimate high-frequency volatility, which they call the ACD-ICV method. They find that ACD-ICV outperforms many version of RV methods in Monte Carlo exercises. Bauwens and Veredas (2004) allow conditional duration to be random and apply the model to trade durations, price durations and volume durations. Their approach is close to a SV model although they do not directly specify or measure volatility.

Cho and Frees (1988) are the first to use passage times of Brownian Motion to estimate volatility, assuming it is constant. Andersen, Dobrev, and Schaumburg (2009) introduce a family of nonparametric volatility estimation using different types of passages times. Their method is a natural dual approach to realized volatility: both are nonparametric, assume volatility is stochastic and focus on estimating integrated volatility over longer periods, usually a day. Their duration-based estimator is robust to jumps and compares favorably to many robust RV type estimators.

This paper proposes a class of stochastic price duration models to estimate high-frequency volatility parametrically. In the baseline model which we call *SPD0*, logarithmic volatility follows an Ornstein-Ulenbeck(OU) process. The OU process is mean-reverting and it leads to an AR(1) process when discretized. The *SPD0* model employs SV models directly in the domain of duration-based estimator.

Interesting extensions to the baseline model can be obtained by incorporating additional information. In particular, we consider trade durations. The asymmetric information models by Easley and O'Hara (1987) suggest that trades durations have an interdependent relationship with volatility. Specifically, since a short trade duration suggests information events and an increased number of informed traders, it tends to be followed by high volatility. On the other hand, lack of trades, or long trade durations are associated with lack of information events and hence lower volatility. Empirical studies also support the impact of duration on volatility.

We model volatility and trade duration using the stochastic volatility and stochastic duration (SVSD) model in Pelletier and Zheng (2012) and Wei and Pelletier (2013). The logarithmic volatility and conditional duration are assumed to follow a bivariate OU process to accommodate their interdependence. We call this model *SPD1.*

A duration-based volatility estimator also faces the challenges from market microstructure noise. The solution is the same as the return-based methods: sample sparsely so the variance of microstructure noise is small compared to volatility. The difference is the sampling scheme: the return-based approach samples at calendar time (e.g. every 5 minutes) or tick time (e.g. every 100 trades), while the duration-based estimator samples at points when the logarithmic price crosses the given threshold; decreasing sampling frequency is achieved by increasing the threshold. In this paper, we choose threshold such that the number of sampling points is comparable to a 5-minute RV estimator.

The sampling scheme renders the first benefit of using price durations over returns: since high volatility results in short price duration, we are sampling more often when the spot volatility is high, and less often when spot volatility is low. Hence, the ratio of noise variance over the volatility integrated over the sampling period is kept relatively flat. Also, if one is interested in the integrated volatility over a day, more points in the realm of high volatility would provide a better approximation to the integration.

The second benefit of using price duration is that it is robust to the discreteness of price. In an ideal world, prices are observed continuously. In the real world, the minimal price change is determined by the tick size, which has been \$0.01 since 2000. Cho and Frees (1988) compare the duration-based approach with the return-based approach in the presence of price discreteness, and they show that low-priced stocks suffer the most from price discreteness and have the most to gain from using duration-based estimators. Intuitively, if the price of stock is \$1, the smallest change of return one can observe is 1% while for a \$100 stock, the smallest increment for return is 0.01%. Price discreteness results in zero returns and complicates estimation for high frequency volatility. Price duration is naturally robust to price discreteness and it is particular advantageous for low-priced stock.

The benefits of using a parametric approach are threefold. First, we can utilize the persistence of volatility. Second, we can estimate intraday spot volatility while the non-parametric approach usually focuses on integrated volatility in a day. Third, we can extend the model to incorporate additional information, such as trade durations.

The rest of this chapter is organized as follows: Section 2 describes the model specification. Section 3 discusses the estimation procedure and conducts simulation studies. Section 4 presents empirical results. Section 5 concludes.

2 Model Specification

2.1 Stochastic Price Duration

We start by assuming that the logarithmic asset price y_t solves the following stochastic differential equations:

$$\begin{aligned} dy_t &= \sqrt{V_t} dW_t^y \\ d\log V_t &= -\kappa^v(\log V_t - \mu^v)dt + \sigma_v dW_t^v, \end{aligned} \tag{1}$$

where V_t is the latent instantaneous variance. W_t^y and W_t^v denote standard Brownian motions. For simplicity, we assume that W_t^y and W_t^v are independent, i.e., there is no leverage effect. From equation (1), we know that the logarithmic volatility follows the OU process:

$$\log V_t = (1 - e^{-\kappa^v(t-s)})\mu^v + e^{-\kappa^v(t-s)} \log V_s + \sigma_v \int_s^t e^{-\kappa^v(t-s)} dW_s^v, \tag{2}$$

where $t > s$. The long-run mean of this process equals μ^v. The parameter κ^v and σ_v describe the persistence and variability of the process, respectively. The long-run variance of logarithmic volatility is given by $\sigma_v^2/2\kappa^v$.

We use price durations to discretize the above process. Price duration is the time it takes for y_t to change by a given amount δ, also called the price threshold. Specifically, if τ_{i+1} is the $i+1$th price duration, $\tau_{i+1} = \inf\{t > 0 | |y_{t_i+t} - y_{t_i}| \geq \delta\}$, where t_i denotes the time when y_t crosses the threshold for the ith time. The sequence $\{t_i\}_{i=0}^N$ partitions the time line $[t_0, t_N]$ into N intervals, while each interval corresponds to a price duration, i.e., $\tau_{i+1} = t_{i+1} - t_i$.

To obtain the distribution of τ_{i+1}, we assume that volatility is constant within each price duration. In other words, we approximate volatility by a piecewise constant process, while the instantaneous volatility in the interval $[t_i, t_{i+1}]$ equals to V_i, the volatility at the left end point of the interval. From the passage time theory for Brownian motion, τ_{i+1} can be written as a function of the price threshold δ and V_i multiplied by a random variable η_{i+1}, see for example Andersen, Dobrev, and Schaumburg (2009). Specifically:

$$\tau_{i+1} = \frac{\delta^2}{V_i}\eta_{i+1}. \tag{3}$$

The random variable η is the price duration when volatility and price threshold are both equal to 1. In passage time theory, η is also referred to as the first exit time, since it measures the time it takes for a standard Brownian motion to exit the band $[-1, 1]$. The distribution of η is given by

$$p(\eta) = \sum_{k=-\infty}^{\infty} \frac{2(1+4k)}{\sqrt{2\pi}\eta^{3/2}} e^{-\frac{(1+4k)^2}{2\eta}}. \tag{4}$$

From equation (2), we get the discretized logarithmic volatility using price durations:

$$\log V_{i+1} = \left(1 - e^{-\kappa^v \tau_{i+1}}\right)\mu^v + e^{(-\kappa^v \tau_{i+1})}\log V_i + u^v_{i+1} \tag{5}$$

where

$$u^v_{i+1} \sim N\left(0, \frac{\sigma_v^2}{2\kappa_v}\left(1 - e^{(-2\kappa^v \tau_{i+1})}\right)\right).$$

Equation (3) and (5) form the discretized baseline model *SPD0*. It is a non-linear non-Gaussian state space model where (3) is the observation equation and (5) is the evolution equation. In the baseline model, we do not consider information from other observables such as number of trades and volume in each price duration. The number of trades is particularly interesting since it reveals the trade durations, which is interdependent with volatility as suggested by the market microstructure theory. We introduce trade durations in the next subsection.

2.2 Stochastic Trade Duration

The trade duration D_{j+1} is defined as the time interval between a trade that occurred at t_j and the next trade at t_{j+1}. Let λ_{t_j} denote the conditional expectation of D_{j+1} given the information set available at t_j, $E(D_{j+1}|I_{t_j}) = \lambda_{t_j}$. We assume that trade durations are exponentially distributed given the conditional duration λ_{t_j}. Hence, D_{j+1} is equal to λ_{t_j} multiplied by an i.i.d random variable with exponential distribution, i.e., $D_{j+1} = \lambda_{t_j} e_{j+1}$. The conditional duration λ_t can vary over time and gives rise to interesting dynamics in trade durations.

Suppose that N trades happened in a time interval with length τ, and we are interested in the distribution of τ given N and λ_t[3]. For simplicity, we assume that the conditional duration λ_t is constant within the time interval. In this case, each trade duration follows an exponential distribution with scale parameter λ_t, and τ is the sum of N exponentially distributed variables. The distribution of τ is given by a gamma distribution with shape parameter N and scale parameter λ_t, $\tau \sim \text{Gamma}(N, \lambda_t)$. We can also look at the distribution of the average duration, $d^a = \tau/N$. Applying the change of variable formula we have $d^a \sim \text{Gamma}(N, \lambda_t/N)$. We can use the scaling property of the gamma distribution to write d^a as λ_t multiplied by a random variable with a $\text{Gamma}(N, 1/N)$ distribution.

In general, if we observe N_{i+1} trades in the time interval $[t_i, t_{i+1}]$ with $\tau_{i+1} = t_{i+1} - t_i$,

[3] We can also use the distribution of N given τ.

the average trade duration $d^a_{i+1} = \tau_{i+1}/N_{i+1}$ can be written as

$$d^a_{i+1} = \lambda_i e_{i+1}, \tag{6}$$

where $e_{i+1} \sim \text{Gamma}(N_{i+1}, 1/N_{i+1})$, and λ_i is the conditional duration at the beginning of the interval. It is easily seen that $E(d^a_{i+1}|I_{t_i}) = \lambda_i$.

2.3 Modeling Price Durations and Trade Durations Jointly

To create persistence and interdependence between volatility and trade duration, we model the logarithm of λ_t and V_t using a bivariate OU process (see Wei and Pelletier, 2013, for properties of this process). Let $x_t = (\log(V_t), \log(\lambda_t))'$, x_t solves:

$$dx_t = -\Psi(x_t - \mu^x)dt + S_x dW^x_t, \tag{7}$$

where Ψ is a 2×2 matrix that measures the mean reversion and dependence between conditional duration and volatility. The process mean reverts to μ^x, the diffusive long-run mean. S_x measures the variation of the logarithmic volatility and the logarithmic duration, and $S_x = \text{diag}(\sigma_v, \sigma_\lambda)$. W^x_t is a Brownian motion in $\mathcal{R}^2$ with $dW^v_t dW^\lambda_t = \rho dt$, where ρ is the instantaneous correlation. The instantaneous covariance matrix is given by

$$\Sigma_x = S_x \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} S_x = \begin{pmatrix} \sigma_v^2 & \rho\sigma_v\sigma_\lambda \\ \rho\sigma_v\sigma_\lambda & \sigma_\lambda^2 \end{pmatrix}.$$

The observables for this model are price duration $\{\tau_i\}_{i=1}^N$ and average duration $\{d^a_i\}_{i=1}^N$. We discretized the bivariate OU process using price durations. As before, we assume that volatility and conditional durations are constant within each price duration. The discretized model *SPD1* is a non-linear and non-Gaussian state space model. The observation equations for *SPD1* are

$$\begin{aligned} \tau_{i+1} &= \frac{\delta^2}{V_i}\eta_{i+1} \\ d^a_{i+1} &= \lambda_i e_{i+1}, \end{aligned} \tag{8}$$

where the distribution of η_{i+1} is given in (4) and $e_{i+1} \sim \text{Gamma}(N_{i+1}, 1/N_{i+1})$. The evolution equation is

$$x_{i+1} = (I_2 - e^{-\Psi\tau_{i+1}})\mu^x + e^{-\Psi\tau_{i+1}}x_i + u_{i+1}, \tag{9}$$

where

$$\begin{aligned} &u_{i+1} \sim N(0, \Sigma_{i+1}) \\ &vec(\Sigma_{i+1}) = (\Psi \oplus \Psi)^{-1}(I_2 - e^{-(\Psi\oplus\Psi)\tau_{i+1}})vec(\Sigma_x). \end{aligned}$$

3 Estimation Procedure and Simulation Studies

3.1 Linear State Space Representation

The inference for models with stochastic volatility or stochastic conditional duration is nontrivial since the evaluation of the likelihood involves integrating out the latent variables. To avoid high dimensional integration, we adopt the quasi-maximum likelihood estimation (QMLE) method that is popular in return-based SV models (see e.g. Harvey, Ruiz, and Shephard, 1994 and Ruiz, 1994). The idea of QMLE is to approximate the non-linear non-gaussian state space model by a linear and gaussian one, and use the Kalman filter to obtain the likelihood. There also exists inference methods that evaluate the exact likelihood, such as simulated maximum likelihood (Danielsson, 1994) or Markov Chain Monte Carlo (Jacquier, Polson, and Rossi, 1994 and Kim, Shephard, and Chib, 1998). However, these methods are computationally intensive, and it is difficult to estimate data in a long period of time given the sample size of high-frequency data. Also, as we will demonstrate later, the approximation error in the QMLE method is less severe in duration-based models than return-based models.

To apply QMLE to the baseline model *SPD0*, we start by taking the logarithm of equation (3):

$$\log \tau_{i+1} = 2 \log \delta - \log V_i + \log \eta_{i+1}, \tag{10}$$

and approximate $\log \eta_{i+1}$ by a normally distributed variable that has the same mean and variance. Equation (10) and (5) form the linear state space representation for *SPD0,* so we can use the Kalman filter to get parameter estimates and smoothed volatility estimates. See de Jong (1989) for the filtering and smoothing procedure with time-varying coefficients.

Parameter estimates yielded by QMLE are consistent and asymptotically normally distributed. The efficiency of the estimator depends on the approximation error; if the true distribution is far from normal, the estimator could be highly inefficient. Return-based estimation requires approximating the logarithm of a chi-squared distribution by a normal distribution, whereas our model approximates the logarithm of price durations as normal. Figure (1) plots the true distribution versus a normal distribution with the same mean and variance for both logarithmic squared returns and logarithmic price durations. As can be seen, the logarithm of price duration is better approximated by the normal distribution[4]. Hence, for our duration-based models, we gain computational speed from

[4] This feature is also shared by range-base estimators, see Alizadeh, Brandt, and Diebold (2002).

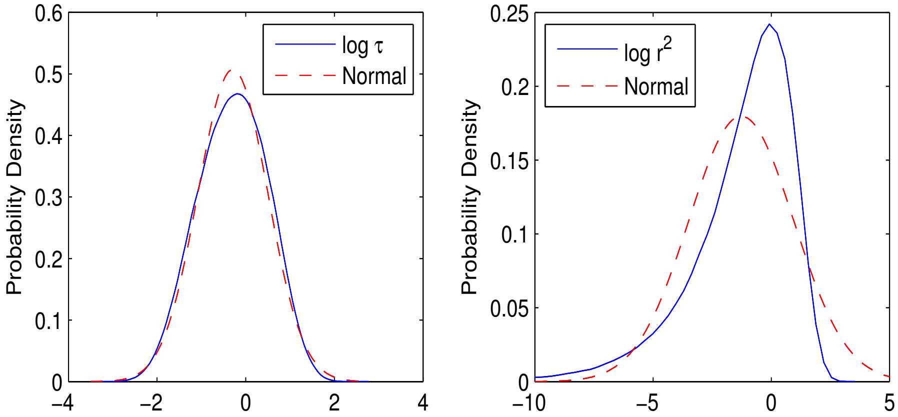

Figure 1: PDF of the true distributions versus their normal approximations. The left panel plots the distribution of logarithmic price durations versus a normal distribution with the same mean and variance. The right panel plots the distribution of squared returns versus its normal approximation.

using QMLE without much loss of efficiency.

If the asset prices have a jump component, as suggested by much empirical work in the literature, the true distribution of price durations would differ. However, price durations have some natural robustness to jumps as demonstrated by Andersen, Dobrev, and Schaumburg (2009) and Tse and Yang (2012). Jumps in the price process might shorten the price duration, but the amount by which the price exceeds the threshold does not directly impact the estimation. Another complication comes from time discreteness: we do not observe price continuously in time, so the actual price change is usually slightly larger than the price threshold δ. This issue can be mitigated if we replace δ by the average actual price change in the MLE. We leave the exact solution to these issues to future work.

We estimate the *SPD1* model using QMLE as well. To linearize the average trade durations, we take the logarithm of equation (6) and approximate $\log e_{i+1}$ by a normal distribution. Since e_{i+1} is distributed as Gamma$(N_{i+1}, 1/N_{i+1})$, the mean and variance of $\log e_{i+1}$ are given by $\psi(N_{i+1}) - \log(N_{i+1})$ and $\psi_1(N_{i+1})$ respectively, where $\psi(x)$ denotes the digamma function and $\psi_1(x)$ denotes the trigamma function. Finally, we have

$$\begin{pmatrix} \log \tau_{i+1} \\ \log d^a_{i+1} \end{pmatrix} = \begin{pmatrix} 2\log\delta + E(\log\eta) \\ E(\log e_{i+1}) \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} x_i + w_{i+1}, \qquad (11)$$

where

$$w_{i+1} \sim N\left(0, \begin{pmatrix} \mathrm{Var}(\log\eta) & 0 \\ 0 & \psi_1(N_{i+1}) \end{pmatrix}\right).$$

Equation (11) and (9) form the linear state space representation of *SPD1* model.

An important issue in applications is to infer the stochastic volatility. We obtain volatility estimates from the smoothed latent variables. Let $x_{i|N}$ and $P_{i|N}$ denote the projection of x_i on all observations and its mean squared error, i.e., $x_{i|N} = E(x_i|\mathcal{F}_N)$ and $P_{i|N} = \mathrm{MSE}(x_{i|N})$, the smoothed estimate for V_i is obtained from the upper left element of $\exp(x_{i|N} + P_{i|N}/2)$.

3.2 Simulations without Microstructure Noise

We perform simulation studies to illustrate the potential gain from using additional information from trade durations (or loss from not using trade durations). We generate logarithmic price and trade durations assuming that conditional duration and volatility are interdependent. Specifically, we use the following parameter value: $(\Psi_{11}, \Psi_{12}, \Psi_{21}, \Psi_{22}) = (0.01, 0.01, 0.02, 0.03)$, $\mu^x = (-18.8, 0.5)'$ and $(\sigma_v, \sigma_\lambda, \rho) = (0.026,\ 0.088,\ -0.5)$. The parameters are chosen such that the annualized volatility is targeted at 20%, and trades happen every 1.6 seconds on average. We then obtain price durations and average trade durations by setting the price threshold to 0.001, which corresponds to approximately 0.1% change in the price.

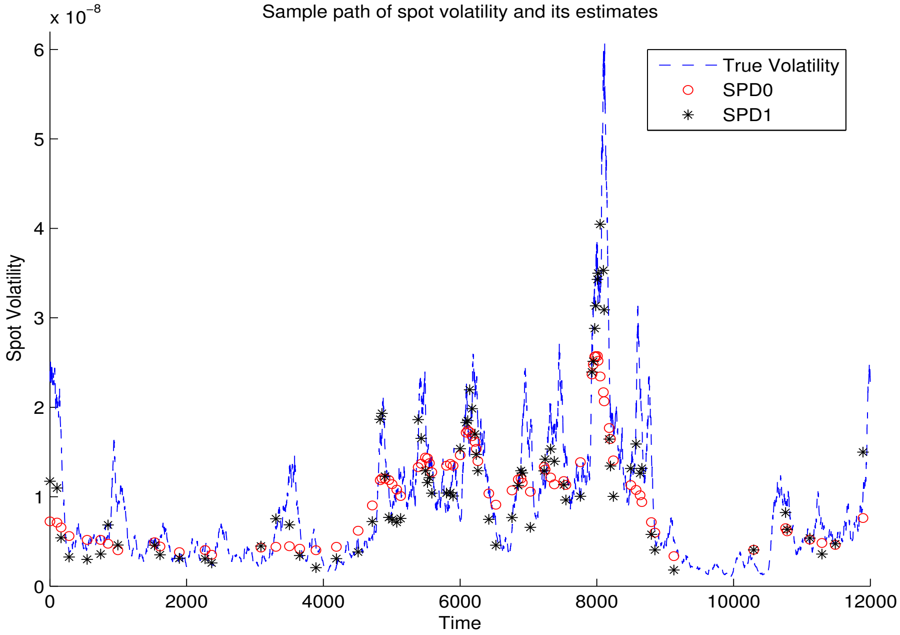

Figure 2: True volatility versus the estimated volatility from *SPD0* and *SPD1*.

Figure 2 presents an example of the true spot volatility and its estimates. Several comments can be made regarding this figure. First, we are estimating more points when volatility is high, and less points when volatility is low. Second, the volatility estimated from both *SPD0* and *SPD1* models are able to capture the main dynamics in spot volatility. Third, by utilizing trade durations, the *SPD1* model outperforms *SPD0* in the sense that estimated volatility from *SPD1* tracks the true volatility more closely. We use the root mean squared error (RMSE) to quantity the difference. The RMSE for each model is computed by RMSE $= \sqrt{\sum_{i=1}^{N}(V_i - \hat{V}_i)^2/N}$, where V_i and $\hat{V}_i$ denote the true and estimated spot variance at time t_i, respectively. The RMSE from the baseline model *SPD0* is 32% higher than the RMSE from the *SPD1* model.

We also plot the true logarithmic conditional duration versus its estimates from *SPD1* model in Figure 3. The estimates trace changes in the true conditional duration although they do not fully capture the rapid fluctuations since we are using average trade durations.

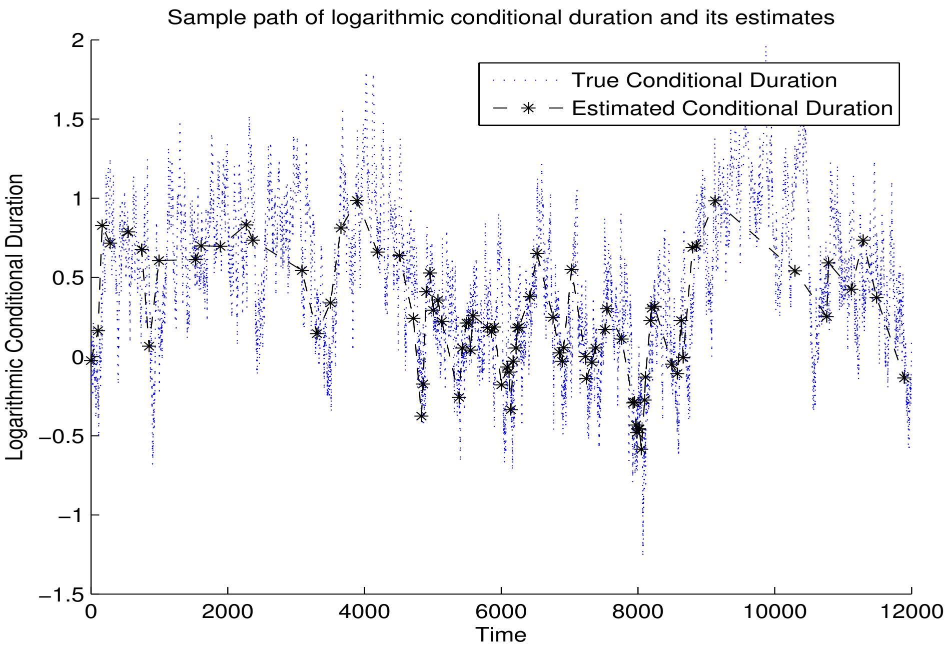

Figure 3: Logarithmic conditional duration and its estimates from *SPD1*.

3.3 Simulation with Microstructure Noise

We compare the performance of SPD0 and SPD1 models to the popular realized volatility approach in the presence of market microstructure noises. Suppose that the observed logarithmic price is contaminated by i.i.d. microstructure noises,

$$y_i^o = y_i + m_i, \tag{12}$$

where $m_i \sim N(0, \sigma_m^2)$. We assume that the true prices are generated from the same stochastic process as in the last subsection. The observed y_i^o are generated with a different Noise-to-Signal Ratio (NSR). Here we define NSR $= \sigma_m / V^{\text{mean}}$, where V^{mean} is the long run mean of spot volatility. We set NSR to (0.25, 1, 1.5), representing low, median and high noise levels.

We conduct 200 simulations, while each simulation consists of data that represents one trading day (6.5 hours or $23,400$ seconds). Realized volatility is then computed by the sum of squared returns within a day, $RV_t = \sum_{j=1}^{1/\Delta} r_{t+j\Delta}^2$. Theoretically, RV_t converges in probability to the integrated volatility over the day, IV_t, when sampling frequency goes to infinity. In the Monte Carlo experiment, we choose the sampling frequency according to the noise levels. Specifically, we sample every 3, 4 or 5 minutes for the low, median or high noise levels. For price durations, we choose price threshold such that the duration-based approach has the same number of observations as the realized volatility approach. In other words, the average price duration is calibrated to 3, 4 or 5 minutes for the low, median or high noise levels. We then estimate the spot volatility from *SPD0* and *SPD1* model, and compute the integrated volatility in a given day by $\widehat{IV_t} = \sum_{i \in day(t)} V_i \tau_{i+1}$. We use RMSE to compare the performance of the estimated IV[5], with RMSE $= \sqrt{\sum_{t=1}^{T}(\widehat{IV}_t - IV_t)^2/T}$.

Table 1 reports the Monte Carlo results for the estimated IV from the *SPD1* model, the *SPD0* model, and the RV approach. It can be seen that both *SPD* models outperforms the realized volatility across different noise levels. Also the *SPD1* models performs better than the *SPD0* model, and the gain increases when NSR increases. This is as expected since *SPD0* does not utilize trade durations, and when NSR increases, prices are more contaminated while trade durations are not affected.

4 Empirical Results

4.1 Data

We apply our model to the milli-second time stamped IBM trade data in the US Equity Data provided by TickData. The sample period is August and September 2011 (44

[5] We could also use RMSE to compare the spot volatility estimates between the *SPD0* and the *SPD1* model as in the previous section, but the RV approach does not provide spot volatility estimates.

Table 1: Monte Carlo results for the estimated IV

	ME	SE	RMSE	Relative RMSE
		NSR = 0.25		
SPD1	-2.88E-06	4.11E-05	1.97E-05	53.9%
SPD0	1.58E-06	4.09E-05	2.01E-05	55.1%
RV	1.71E-06	5.46E-05	3.65E-05	100%
		NSR = 1		
SPD1	1.30E-05	4.45E-05	2.56E-05	60.9%
SPD0	1.87E-05	4.48E-05	2.98E-05	71.0%
RV	4.32E-06	5.69E-05	4.20E-05	100%
		NSR = 1.5		
SPD1	2.60E-05	4.87E-05	3.69E-05	81.1%
SPD0	3.34E-05	4.98E-05	4.33E-05	95.2%
RV	-1.07E-06	5.79E-05	4.55E-05	100%

Notes: ME = mean error, SE = standard deviation of sample estimates. RMSE = root mean squared error. The last column express the RMSE from the *SPD1* and *SPD0* model as a percentage of the RMSE from RV. We conduct 200 Monte Carlo simulations, while each simulation corresponds one trading day.

trading days). We follow the cleaning procedure proposed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009) to filter out the potentially erroneous data. First, entries with a correction indicator other than 0 are deleted. Second, we delete entries with abnormal sales conditions (see the TAQ manual for a complete reference on the correction indicator and sales condition). Third, observations from outside of the normal opening time are omitted. Fourth, we delete entries from the first five minutes after opening to eliminate the price changes due to information accumulated overnight. Last, we treat entries within 0.1 second as one observation and use the mean price to alleviate possible measurement error in the transaction time.

To obtain price durations, we set the price threshold to 0.002 (roughly a 0.2% change in price) so that average price duration is roughly 5 minutes. This results in a total of 3,361 sampling points. Figure 4 illustrates an example of the sampling points in a day. As we can see, the sampling points are more concentrated near the beginning, when

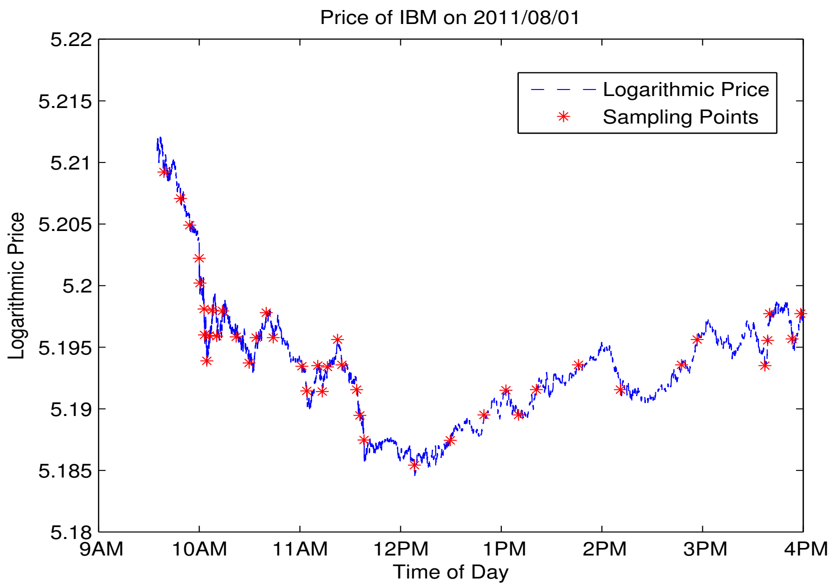

Figure 4: Sampling points in one trading day. We sample when the change in the logarithmic price reaches or exceeds 2%.

Table 2: Summary statistics for IBM in 2011/08/01-2011/09/31

	Mean	Median	Standard Deviation
Price Duration	299	144	465
Number of Trades per price duration	209	132	245
Average Trade Duration	1.25	1.07	0.76

Notes: All units reported are in seconds.

price is changing violently. We then divide each price duration by the number of trades within that price duration to obtain average trade durations. Summary statistics for the observables is given in Table 2. Since trades are occurring frequently (every 1.25 seconds on average), the impact from time discreteness is minimal.

4.2 Diurnal Pattern

Intraday volatility and duration have well known diurnal patterns. Transactions happen more frequently near the opening and closing times, and less frequently during the middle of a day. This deterministic diurnal pattern needs to be accounted for before we specify a stochastic model for the latent variables.

We use a quadratic function[6] to approximate the diurnal pattern and estimate it within the model. The level of the quadratic function is fixed by setting its minimum to 1, otherwise the mean of the latent process becomes unidentifiable. For the *SPD1* model (the procedure for *SPD0* model naturally follows), we adopt the following quadratic functions for volatility and conditional duration:

$$\begin{aligned} g_v(t) &= a_1(t+a_2)^2+1, \\ g_d(t) &= a_3(t+a_4)^2+1, \end{aligned} \tag{13}$$

Letting V_i^* and λ_i^* denote the deseasonalized volatility and conditional duration, we have

$$\begin{aligned} V_i &= V_i^* g_v(t_i), \\ \lambda_i &= \frac{\lambda_i^*}{g_d(t_i)}. \end{aligned} \tag{14}$$

This specification produces the U-shaped patten in volatility and the inverse U-shaped patten in the conditional duration. After considering the diurnal effect, the observation

[6] The choice of a quadratic function is a trade-off between better approximation and less parameters to estimate. The nonparametric estimate in Chapter 1 indicates that a quadratic function describes the main dynamics of the diurnal pattern. Higher order approximation may improve the fit, and we leave that to future work.

equation for *SPD1* model becomes

$$\begin{pmatrix} \log \tau_{i+1} \\ \log d^a_{i+1} \end{pmatrix} = \begin{pmatrix} 2\log\delta + E(\log\eta) - \log g_v(t_i) \\ E(\log e_{i+1}) - \log g_d(t_i) \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} x_i + w_{i+1}, \qquad (15)$$

where $x_i = (\log V^*_i,\, \log\lambda^*_i)'$ and it follows the evolution equation (9).

4.3 Estimation

We estimate the data in the sample period using both the *SPD0* and *SPD1* models. To deal with observations from different trading days, we assume that each day "starts fresh": the latent OU process starts with its long-run mean and variance each day. The parameter estimates from the *SPD0* and *SPD1* models are presented in Table 3. In the *SPD0* model, the parameter estimates indicate an annualized volatility of 25%. In the *SPD1* model, the market microstructure theory from Easley and O'Hara (1987) predicts that high volatility leads to short durations, while short durations have a positive effect on volatility. In our estimate, the impact of volatility on conditional duration is profound, while the effect of conditional duration on volatility and their instantaneous correlation is not statistically significant.

We plot the diurnal patterns estimated from the *SPD1* model in Figure 5. The diurnal pattern indicates that the volatility near the beginning of a trading day is almost 5 times as big as its minimum around noon. The conditional trading durations are less than half as long as the conditional trading durations near the middle of the day.

We compare the parametric *SPD* models to the nonparametric RV approach as well. Figure 6 presents the daily integrated volatility estimated from the 5-minute RV, *SPD0* model and *SPD1* model. The integrated volatility in the parametric models is obtained from the smoothed estimates of spot volatility, $\widehat{IV}_t = \sum_{i\in day(t)} V_i\tau_{i+1}$. As we can see, the

Table 3: Parameter estimates for the *SPD1* and *SPD0* model

	SPD1	std. error	*SPD0*	std. error
Ψ_{11}	3.91E-04	6.58E-03	4.09E-04	8.38E-05
Ψ_{21}	0.049	0.018		
Ψ_{12}	1.11E-10	1.51E-02		
Ψ_{22}	0.115	0.040		
μ^v	-19.293	0.162	-19.103	0.094
μ^d	0.911	0.054		
σ_v	0.025	0.005	0.026	0.003
σ_d	0.129	0.021		
ρ	-0.546	0.561		
a_1	4.14E-08	1.13E-08	2.28E-08	3.99E-09
a_2	-1.07E+04	4.93E+02	-1.29E+04	5.17E+02
a_3	1.56E-08	1.98E-09		
a_4	-1.03E+04	3.10E+02		

Notes: We estimate the *SPD1* and *SPD0* model using milli-second IBM data in 2011/08/01-2011/09/30.

three volatility estimates trace each other quite closely.

5 Conclusion

In this paper we present a new parametric model to estimate stochastic volatility based on price durations. This model has several advantages: first, price durations have some robustness to jumps and market microstructure noise, especially the noise from price discreteness. Second, we utilize the persistence of volatility and we can infer volatility integrated over any period of time. Third, contrary to the ACD-type models, we assume that volatility is stochastic, and we obtain the distribution of price durations from the passage theory for Brownian motions. Last, we can conduct inference easily using QMLE without much loss of efficiency since the logarithmic price duration is better approximated by a normal distribution than the logarithmic squared returns.

We also extend the baseline model *SPD0* to incorporate information from trading

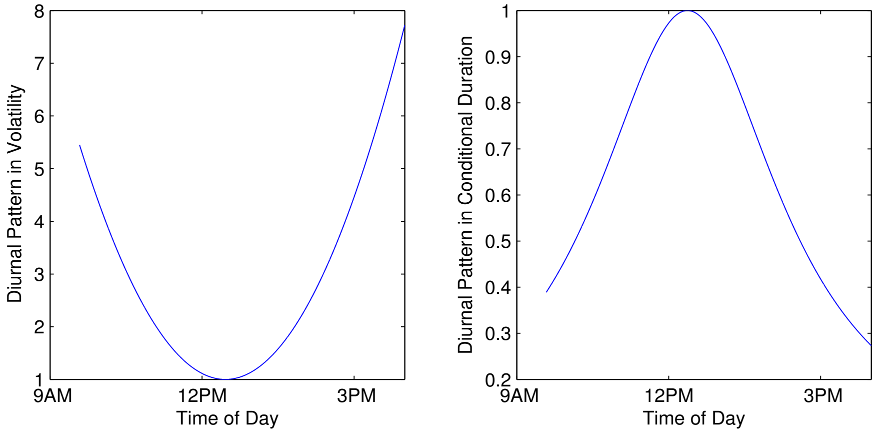

Figure 5: Diurnal patterns in volatility and conditional duration.

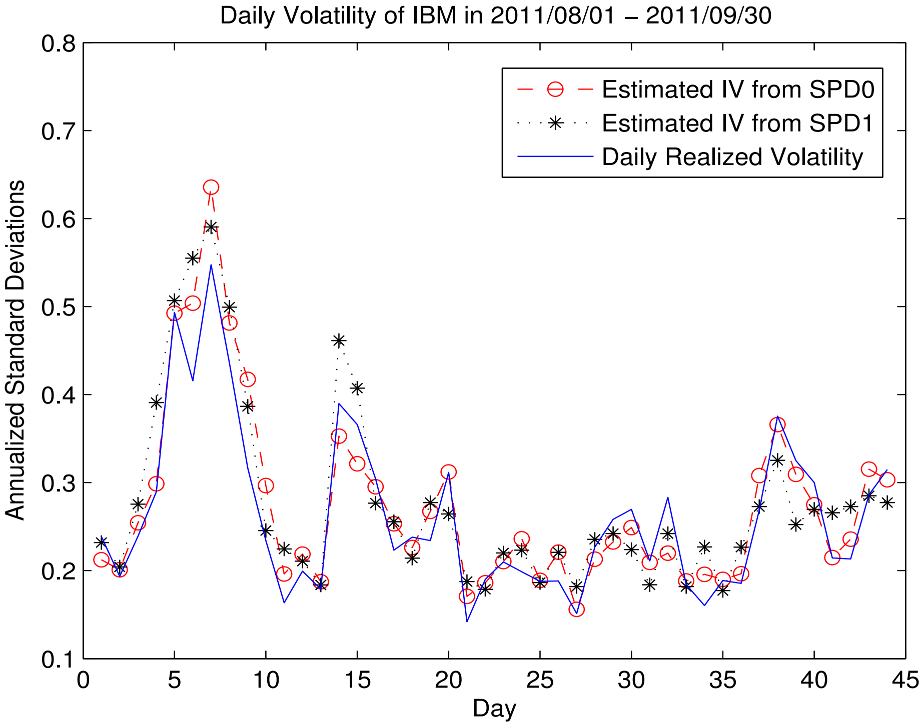

Figure 6: Daily volatility of IBM in 2011/08/01-2011/09/31, 44 days.

durations, as market microstructure theory suggests that trading durations and volatility are interdependent. We call the more sophisticated model *SPD1.* We conduct Monte Carlo studies to demonstrate the performance of the price duration models. We find that *SPD1* outperforms *SPD0* in estimating spot volatility, and both duration-based model performs better than realized volatility in estimating the integrated volatility.

There are several interesting extensions we can explore in this class of price duration models. First, volume is another variable that could influence volatility, and hence could be incorporated into the latent process. Second, we can consider the distribution of price duration if the asset price follows a jump diffusion. Last, we can further investigate the influence of time discreteness on the distribution of price durations, especially for less liquid stocks.

References

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). 'Range-Based Estimation of Stochastic Volatility Models', *The Journal of Finance*, 57(3): 1047–1091.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2009). 'Duration-Based Volatility Estimation', Global COE Hi-Stat Discussion Paper Series gd08-034, Institute of Economic Research, Hitotsubashi University.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). 'Realized kernels in practice: trades and quotes', *Econometrics Journal*, 12(3): C1–C32.

Bauwens, L., and Veredas, D. (2004). 'The stochastic conditional duration model: a latent variable model for the analysis of financial durations', *Journal of Econometrics*, 119(2): 381–412.

Cho, D. C., and Frees, E. W. (1988). 'Estimating the Volatility of Discrete Stock Prices', *Journal of Finance*, 43(2): 451–66.

Danielsson, J. (1994). 'Stochastic volatility in asset prices estimation with simulated maximum likelihood', *Journal of Econometrics*, 64(1-2): 375–400.

de Jong, P. (1989). 'Smoothing and Interpolation with the State-Space Model', *Journal of the American Statistical Association*, 84(408): pp. 1085–1088.

Easley, D., and O'Hara, M. (1987). 'Price, trade size, and information in securities markets', *Journal of Financial Economics*, 19(1): 69–90.

Engle, R. F., and Russell, J. R. (1998). 'Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data', *Econometrica*, 66(5): 1127–1162.

Hansen, P. R., and Lunde, A. (2005). 'A forecast comparison of volatility models: does anything beat a GARCH(1,1)?', *Journal of Applied Econometrics*, 20(7): 873–889.

Harvey, A., Ruiz, E., and Shephard, N. (1994). 'Multivariate Stochastic Variance Models', *Review of Economic Studies*, 61(2): 247–64.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). 'Microstructure noise in the continuous case: The pre-averaging approach', *Stochastic Processes and their Applications*, 119(7): 2249–2276.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). 'Bayesian Analysis of Stochastic Volatility Models', *Journal of Business & Economic Statistics*, 12(4): 371–89.

Kim, S., Shephard, N., and Chib, S. (1998). 'Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models', *Review of Economic Studies*, 65(3): 361–93.

Pelletier, D., and Zheng, H. (2012). 'Joint Modeling of High-Frequency Price and Duration Data', Discussion paper, North Carolina State University.

Ruiz, E. (1994). 'Quasi-maximum likelihood estimation of stochastic volatility models', *Journal of Econometrics*, 63(1): 289 – 306.

Tse, Y. K., and Yang, T. T. (2012). 'Estimation of High-Frequency Volatility: An Autoregressive Conditional Duration Approach', *Journal of Business & Economic Statistics*, 30(4): 533–545.

Wei, W., and Pelletier, D. (2013). 'A Jump Diffusion Model for Volatility and Duration', Discussion paper, North Carolina State University.

Zhang, L., Mykland, P. A., and Ait-Sahalia, Y. (2005). 'A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data', *Journal of the American Statistical Association*, 100: 1394–1411.