Zhou's estimator

RK

August 29, 2014

Abstract

The paper by Bin Zhou, titled "High Frequency Data and Volatility in Foreign-Exchange Rates" is one of the first papers in the finance literature to address the problem of volatility estimation in the presence of market microstructure noise. This document explains the rationale and the math behind the estimate.

Introduction

One of the areas where extensive research has been done, is the area of *volatility estimation*, more so in the last 20 years where high frequency data (HFD) is being increasingly available to researchers. The abundance of data has its flip side; one needs to deal with the microstructure noise in building models. The paper by Zhou contains an estimate for volatility in the presence of microstructure noise. Before one begins understanding the estimate, it is better to get an idea of volatility estimation in the absence of microstructure noise.

Let $X(t)$ be log price process and for simplicity let us assume it to be a driftless brownian motion.

$$dX(t) = \sigma(t)dW(t)$$

If one has infinite data points, we can get an estimate of instantaneous volatility ($\sigma(t)$). Since we are dealing with a discrete ticks, one can approximate the total variance

$$Q = \int_0^T \sigma^2(t)\, dt$$

In a small time interval (t_{j-1}, t_j), we have

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

Since this is an Ito integral, the increment is normally distributed. Hence the square of this increment is a χ^2 distribution with 1 degree of freedom. Given that the quadratic variation of the log price increment is $\langle dX_t, dX_t \rangle = \sigma^2(t)dt$, we can use the following as an estimator for volatility (under the assumption of constant volatility)

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

Alas! World is not so simple. We live in a constantly *varying* volatility. Hence the best we can do is get our hands around *realized variance*

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

If N is large, then the realized variance can be written as

$$RV = Q + \xi, \quad \xi \sim \mathcal{O}(\sigma^4 T^2/N)$$

and hence the volatility estimate is

$$\hat{\sigma}^2_{RV} = RV/T$$

Thus we can use either one of the above estimators. The former is good in the presence of constant volatility (almost never) and the latter is useful when volatility varies in some unknown way. Most of the elementary finance textbooks stop with this definition and do not explore the next obvious issue, "estimation in the presence of microstructure noise".

Rationale behind Zhou's estimate

One of the easiest ways to incorporate noise is to add an additional term to the log price increments

$$dY(t) = \sigma(t)dW(t) + d\epsilon_t$$

where ϵ_t is a realization of an iid process. Note that this model is being introduced to take care of the noise that arises additional to bid-ask bounce. Bid-ask bounce does induce a negative correlation between tick movement but for modeling purpose that can be easily removed by taking quote mid point prices.

Given the above mentioned price evolution, what is the realized volatility ?

$$RV = \sum_{j=1}^{N} (y_j - y_{j-1})^2 = \sum_{j=1}^{N} (x_j + \epsilon_j - x_{j-1} - \epsilon_{j-1})^2$$

Since the model has guassian written all over the terms, RV in this new setting is also gaussian

$$RV = Q + 2N\eta^2 + \xi, \quad Var(\xi) \sim \mathcal{O}(N\eta^4 + N\sigma^4\tau^2)$$

where η is the variance of the noise term. As one can see the above estimate is clearly biased. The more finer the sampling frequency, the more biased is the RV. One easy way to get out of this bias is to sample at a low frequency, let's say every kth tick.

ZHOU'S CONTRIBUTION IS THE CORRECTION TERM :
By considering

$$RV = \sum_{j=1}^{N} (y_j - y_{j-1})^2 + y_j y_{j+1} + y_j y_{j-1}$$

one can see that the additional terms kill the bias arising out of the previous RV definition. Zhou's paper states a theorem that derives the variance of this estimator.The proof is given in one step that leaves the reader to work out the details. I happened to derive the estimator and found that there is a typo in the paper. The problem with reading academic papers, unlike books, is that there is hardly any errata available to check for mistakes in the paper. In any case, I tried simulating some code to check whether my derivation had some chinks. Looks there aren't. Here is the statement of the proof

Theorem 1 *The realized variance via the model*

$$\hat{\sigma}_U^2 = \sum_{j=1}^{N} (y_i^2 + y_i y_{i-1} + y_{i+1} y_i)$$

The variance of the estimate is

$$var(\hat{\sigma}_U^2) = \sigma^4 \left(\frac{6}{n} + 8\frac{\eta^2}{\sigma^2} + 8\frac{n\eta^4}{\sigma^4} \right) - 2\frac{\sigma^4}{n^2} - 4\eta^4$$

I have tried deriving the above vol of vol estimate and I have obtained a slightly different form. Here are the main steps :

The objective is to obtain $var(\hat{\sigma}_U^2)$. Since it is a summation of terms, let us look at an individual term

$$(y_i^2 + y_i y_{i-1} + y_{i+1} y_i) = V_i$$

Since the model considered is $dY(t) = \sigma(t)dW(t) + d\epsilon_t$, one can discretize and obtain

$$y_i = \frac{\sigma}{\sqrt{n}} z_i + \epsilon_i - \epsilon_{i-1}$$

Thus one can rewrite V_i as

$$\left(\frac{\sigma}{\sqrt{n}} z_i + \epsilon_i - \epsilon_{i-1}\right)^2 + \left(\frac{\sigma}{\sqrt{n}} z_i + \epsilon_i - \epsilon_{i-1}\right) \cdot \left(\frac{\sigma}{\sqrt{n}} z_{i-1} + \epsilon_{i-1} - \epsilon_{i-2}\right) + \left(\frac{\sigma}{\sqrt{n}} z_i + \epsilon_i - \epsilon_{i-1}\right) \cdot \left(\frac{\sigma}{\sqrt{n}} z_{i+1} + \epsilon_i - \epsilon_i\right)$$

This simplifies to

$$V_i = \frac{\sigma^2}{n}\left(z_i^2 + z_{i+1} z_i + z_{i-1} z_i\right) + \frac{\sigma}{\sqrt{n}}\left(z_i \epsilon_{i+1} + z_i \epsilon_i - z_i \epsilon_{i-1} - z_i \epsilon_{i-2} + z_{i+1} \epsilon_i - z_{i+1} \epsilon_{i-1} + z_{i-1} \epsilon_i - z_{i-1} \epsilon_{i-1}\right) +$$

$$\epsilon_{i+1}\epsilon_i - \epsilon_{i+1}\epsilon_{i-1} - \epsilon_i \epsilon_{i-2} + \epsilon_{i-1}\epsilon_{i-2}$$

$$\begin{aligned} E(V_i) &= \frac{\sigma^2}{n} \\ Var(V_i) &= \left(\frac{\sigma^4}{n^2}\right)(3+1+1) + \frac{\sigma^2}{n} 8\eta^2 + 4\eta^4 - \left(\frac{\sigma^4}{n^2}\right) \\ &= 4\left(\frac{\sigma^4}{n^2}\right) + \frac{\sigma^2}{n} 8\eta^2 + 4\eta^4 \end{aligned}$$

Consider V_{i+1}

$$V_{i+1} = \frac{\sigma^2}{n}\left(z_{i+1}^2 + z_{i+2} z_{i+1} + z_i z_{i+1}\right) + \frac{\sigma}{\sqrt{n}}\left(z_{i+1}\epsilon_{i+2} + z_{i+1}\epsilon_{i+1} - z_{i+1}\epsilon_i - z_{i+1}\epsilon_{i-1} + z_{i+2}\epsilon_{i+1} - z_{i+2}\epsilon_i + z_i \epsilon_{i+1} - z_i \epsilon_i\right) +$$

$$\epsilon_{i+2}\epsilon_{i-1} - \epsilon_{i+2}\epsilon_i - \epsilon_{i+1}\epsilon_{i-1} + \epsilon_i \epsilon_{i-1}$$

One can compute the covariance between V_i and V_{i+1} (expection for most of the product terms is 0)

$$\begin{aligned} Cov(V_i, V_{i+1}) &= \left(\frac{\sigma^4}{n^2}\right)(1+1) + \eta^4 - \left(\frac{\sigma^4}{n^2}\right) \\ &= \left(\frac{\sigma^4}{n^2}\right) + \eta^4 \end{aligned}$$

Consider V_{i+2}

$$\begin{aligned} V_{i+2} = {} & \frac{\sigma^2}{n}\left(z_{i+2}^2 + z_{i+3} z_{i+2} + z_{i+1} z_{i+2}\right) + \\ & \frac{\sigma}{\sqrt{n}}\left(z_{i+2}\epsilon_{i+3} + z_{i+2}\epsilon_{i+2} - z_{i+2}\epsilon_{i+1} - z_{i+2}\epsilon_i + z_{i+3}\epsilon_{i+2} - z_{i+3}\epsilon_{i+1} + z_{i+1}\epsilon_{i+2} - z_{i+1}\epsilon_{i+1}\right) + \\ & \epsilon_{i+3}\epsilon_i - \epsilon_{i+3}\epsilon_{i+1} - \epsilon_{i+2}\epsilon_i + \epsilon_{i+1}\epsilon_i \end{aligned}$$

One can compute the covariance between V_i and V_{i+2} (expection for most of the product terms is 0)

$$\begin{aligned}Cov(V_i, V_{i+2}) &= \left(\frac{\sigma^4}{n^2}\right)(1) + \eta^4 - \left(\frac{\sigma^4}{n^2}\right) \\ &= \eta^4\end{aligned}$$

It is easy to check for all the random varaibles $V_{i+3}, V_{i+4}, \dots$,

$$Cov(V_i, V_{i+k}) = 0, \quad \forall k \geq 3$$

For notational ease, assume

$$\begin{aligned}K_1 &= 4\left(\frac{\sigma^4}{n^2}\right) + \frac{\sigma^2}{n}8\eta^2 + 4\eta^4 \\ K_2 &= \left(\frac{\sigma^4}{n^2}\right) + \eta^4 \\ K_3 &= \eta^4 \\ K_j &= 0 \quad \forall j \geq 4\end{aligned}$$

The distrubution of Random vector $\mathbf{V}$ is a multivariate normal and its covariance matrix is

$$\Sigma = \begin{pmatrix} K_1 & K_2 & K_3 & 0 & \dots & \dots & 0 \\ K_2 & K_1 & K_2 & K_3 & \dots & \dots & 0 \\ K_3 & K_2 & K_1 & K_2 & K_3 \dots & \dots & 0 \\ \vdots \\ \vdots \\ 0 & \dots & \dots & K_3 & K_2 & K_1 & K_2 \\ 0 & \dots & \dots & \dots & K_3 & K_2 & K_1 \end{pmatrix}$$

The variance of Zhou's estimate is

$$var(\hat{\sigma}_U^2) = \mathbf{1}^T \Sigma \mathbf{1}$$

Since it is banded matrix, the variance works out to

$$\begin{aligned}var(\hat{\sigma}_U^2) &= nK_1 + 2(n-1)K_2 + 2(n-2)K_3 \\ &= n4\left(\frac{\sigma^4}{n^2}\right) + \frac{\sigma^2}{n}8\eta^2 + 4\eta^4 + 2(n-1)\left(\frac{\sigma^4}{n^2}\right) + \eta^4 + 2(n-2)\eta^4 \\ &= \sigma^4\left(\frac{6}{n} + 8\frac{\eta^2}{\sigma^2} + 8\frac{n\eta^4}{\sigma^4}\right) - 2\frac{\sigma^4}{n^2} - 6\eta^4\end{aligned}$$

Clearly the last term in the above expression is different from that mentioned in the paper. I have obtained $-6\eta^4$ instead of $-4\eta^4$. This being an old paper, I should search google for the author's contact and then somehow try to get a clarification of the formula in the paper.One thing to note is that the difference is in the coefficient of square of noise variance and hence is going to be negligible in the final volatilty of volatility estimate.

In any case I did the next best thing I could. I have written some basic `R` code to simulate data that can be used to compute intraday volatility assuming 6 hrs of trading time in a day and samples taken at every 5 minutes($N = 12 * 6 = 72$). Have assumed some arbitrary values for stock variance and microstructure noise variance. My objective was to check whether Zhou's closed form solution verifies with the simulated data.

Simulate data

```
sigma.stock <- 0.3/sqrt(252)
sigma.noise <- 0.1*sigma.stock
res <- replicate(10000,{
  N           <- 72
  n           <- N+2
  Z           <- sigma.stock*rnorm(n+1)/sqrt(N)
  eps         <- rnorm(n+1,0,sigma.noise)
  Y           <- numeric(N)
  for(i in 3:n){
    t1        <- Z[i] + eps[i]-eps[i-1]
    t2        <- Z[i-1] + eps[i-1]-eps[i-2]
    t3        <- Z[i+1] + eps[i+1]-eps[i]
    Y[i-2]    <- t1*(t1+t2+t3)
  }
  Y
})
res <- t(res)
```

Estimate using K1,K2,K3,K4

```
vol.vol <- function(N, sigma.stock, sigma.noise){
  K1 <- 4*sigma.stock^4/N^2+ 8*sigma.noise^2*sigma.stock^2/N + 4 *sigma.noise^4
  K2 <- sigma.stock^4/N^2 + sigma.noise^4
  K3 <- sigma.noise^4
  K4 <- 0
  N*K1 + 2*((N-1)*(K2) + max(0,(N-2)*K3)+ max(0,(N-3)*(N-2)/2*K4))
}
```

Estimate using Zhou's formula

```
vol.vol.paper <- function(N, sigma.stock, sigma.noise){
  sigma.stock^4*
    ( 6/N + 8*sigma.noise^2/sigma.stock^2 + 8* N*sigma.noise^4/sigma.stock^4) -
    2*sigma.stock^4/N^2 - 4*sigma.noise^4
}
```

```
N              <- 72
sim.estimate   <- var(rowSums(res))
zhou.estimate <- vol.vol.paper(N,sigma.stock,sigma.noise)
my.estimate    <- vol.vol(N,sigma.stock,sigma.noise)
results <- 100* data.frame(simulated = sqrt(zhou.estimate)*72,
zhou = sqrt(my.estimate)*72, corrected = sqrt(sim.estimate)*72)
```

The following are daily volatility estimates from the simulated dataset, Zhou’s closed form and the corrected form that I have derived in the paper

simulated	zhou	corrected
1.21	1.21	1.20

As one can see, all three are practically same. Hence this serves as an additional check to the formula derived in this document and in the paper.

Summary of Zhou’s paper

I will try to briefly summarize the main points of Bin Zhou’s paper, “High Frequency Data and Volatility in Foreign-Exchange Rates”

The paper analyzes HFD data for DEM/USD, JPY/DEM, JPY/USD. The author takes the bid prices and finds that there is over 40% negative autocorrelation in the tick by tick return. Since there is no bid ask bounce in the data, the author investigates this finding by building a model for the evolution of log stock price by adding an additional term to include microstructure noise. Using this model the author derives an unbiased estimator for realized volatility

$$RV = \sum_{j=1}^{N} (y_j - y_{j-1})^2 + y_j y_{j+1} + y_j y_{j-1}$$

The paper derives the variance of the estimate in terms of sampling rate and find the optimum sampling rate that gives the minimium variance estimate. The optimal number of observations n turns out to be

$$n_{opt} = \frac{\sqrt{3}\sigma^2}{2\eta^2}$$

The author uses the n_{opt} and resamples at a rate that is closer to n_{opt}. Thus various k tick aggregated return are computed over a grid and sensible k is chosen for each of the analyzed dataset.

The author uses the estimate of volatility in standardized daily returns, hourly returns. Using QQ plots, one can clearly see that QQ plots of standardized returns are close to the standard gaussian assumption. This simple estimator can be quickly implemented in any trading system.