The University of Chicago
Department of Statistics

Are Volatility Estimators Robust with Respect to Modeling Assumptions?

Yingying Li and Per A. Mykland

Departments of Statistics
The University of Chicago
Chicago, Illinois 60637

Are Volatility Estimators Robust with Respect to Modeling Assumptions?

Yingying Li* and Per A. Mykland**

November 8, 2006

Department of Statistics, The University of Chicago, Chicago, IL 60637.
Email: *yyli@galton.uchicago.edu. **mykland@galton.uchicago.edu.

Abstract

We consider microstructure as an arbitrary contamination of the underlying latent securities price, through a Markov kernel Q. Special cases include additive error, rounding, and combinations thereof. Our main result is that, subject to smoothness conditions, the two scales realized volatility (TSRV) is robust to the form of contamination Q. To push the limits of our result, we show what happens for some models involving rounding (which is not, of course, smooth) and see in this situation how the robustness deteriorates with decreasing smoothness. Our conclusion is that under reasonable smoothness, one does not need to consider too closely how the microstructure is formed, while if severe non-smoothness is suspected, one needs to pay attention to the precise structure and also to what use the estimator of volatility will be put.

Keywords: Bias Correction; Local Time; Market Microstructure; Martingale; Measurement Error; Robustness; Realized Volatility; Subsampling; Two Scales Realized Volatility (TSRV).

1   INTRODUCTION

Recent years has seen an explosion of literature on the problem of estimating integrated volatility and similar objects with the help of high frequency data. For a sample of recent literature, see Hull & White (1987), Jacod & Protter (1998), Gallant, et al. (1999), Chernov & Ghysels (2000), Gloter (2000), Andersen, et al. (2001), Dacorogna, et al. (2001), Barndorff-Nielsen & Shephard (2002), and Mykland & Zhang (2006) among others. An important realization has been that log prices do not appear to be semimartingales, but rather like semimartingales observed with error. Main hypotheses proposed in the literature is that this error occurs by rounding (Delattre & Jacod (1997), Jacod (1996), Zeng (2003)) or by additive error (Zhou (1996), Zhang, et al. (2005), Zhang (2006), Aït-Sahalia, et al. (2005), Bandi & Russell (2003), Hansen & Lunde (2006)). More complex (and descriptive) models for microstructure are also available, see for example Hasbrouck (1996) and, coming from a very different perspective, by Farmer, et al. (2005).

The multiplicity of ways in which errors can be modeled raises the question of how sensitive inference is to modeling assumptions. This is the topic of this paper.

We shall be making the assumption that there is a latent log price process X_t which is a continuous semimartingale on the form

$$dX_t = \mu_t dt + \sigma_t dB_t, \tag{1}$$

where μ_t and σ_t are continuous random processes, and B_t is a Brownian motion. This is also called an Itô process. Transactions at times $0 = t_0 < t_1 < ... < t_n = T$ give rise to log prices Y_{t_i} which are contaminated versions of X_{t_i} as follows. We suppose that there is a family $Q(x, dy)$ of conditional distributions so that given X_{t_i}, the law of Y_{t_i} is

$$P(Y_{t_i} \leq y \mid X \text{ process}) = P(Y_{t_i} \leq y \mid X_{t_i}) = Q(X_{t_i}, y). \tag{2}$$

In other words, Y_{t_i} is distributed around X_{t_i} in a way that only depends on the latter. We also assume that $Y_{t_0}, ..., Y_{t_n}$ are conditionally independent given the X process.

A simple example of such contamination Q is additive error on the log scale. If $Y = X + \epsilon$ where ϵ has density g and is independent of X, then

$$Q(x, dy) = g(y - x)dy. \tag{3}$$

Another example is rounding or truncation. In this case, the probability distribution $Q(x, dy)$ represents a nonrandom distortion of x. We shall look at yet another form of contamination in Section 3. In that case, the distortion is a combination of additive error and rounding.

This paper has two pieces of news, one good and one bad. We shall see that for reasonable types of contaminations Q, one can act as if the error was simply of additive type, and we shall see that the two scales realized volatility (TSRV) of Zhang, et al. (2005) is substantially robust to arbitrary contamination. This is our plan for Section 2.

There are, however, cases when one has to exercise care. We shall see one such case in Section 3, where we shall show that it is not always quite clear what is meant by volatility, and one has to consider carefully what quantity one actually wishes to estimate. This occurs in cases involving rounding.

We are mainly using TSRV by way of example of a volatility estimator, and believe that similar conclusions will apply to, for example, the multi-scale realized volatility (MSRV) of Zhang (2006). With caveats about additional bias and variance, similar conclusions will also apply to traditional realized volatility (RV).

2 ROBUSTNESS and SMOOTHNESS of CONTAMINATION

2.1 Setup

Suppose the latent log price process X follows (1). Let Y be the logarithm of the transaction price, which is observed at times $0 = t_0 < t_1 < ... < t_n = T$. We assume that at these sampling times, Y

is related to the latent log price process X through (2). Let $f(X_t)$ be the conditional expectation of Y_t given the X process:

$$f(X_t) = E_Q(Y_t|X_t). \tag{4}$$

We shall assume that

$$f(x) \text{ is twice continuously differentiable.} \tag{5}$$

Note that under the assumption (5), $f(X_t)$ is a continuous semimartingale.

Definition 1. *For two generic processes $Z^{(1)}$ and $Z^{(2)}$ and for an arbitrary grid $\mathcal{H} = \{s_0, s_1, s_2, \cdots, s_m\}$ of points in the interval $[0, T]$, define*

$$[Z^{(1)}, Z^{(2)}]_T^{\mathcal{H}} = \sum_{j=1}^{m} (Z_{s_j}^{(1)} - Z_{s_{j-1}}^{(1)})(Z_{s_j}^{(2)} - Z_{s_{j-1}}^{(2)}).$$

Definition 2. *If Z is a continuous semimartingale, its quadratic variation $\langle Z, Z \rangle_T$ is defined as the limit of $[Z, Z]_T^{\mathcal{H}_m}$ if $\mathcal{H}_m$ becomes dense in $[0, T]$ as $m \to \infty$. The quadratic variation is also known as the (integrated) volatility of Z for the time period [0,T].*

The above definition gives a well defined limit $\langle Z, Z \rangle_T$ (independent of the sequence $\mathcal{H}_m$) in view of Theorem 4.47-4.48 (page 52) of Jacod & Shiryaev (2003).

A central problem is that we have two continuous semimartingale processes, X and $f(X)$, which produce two volatilities

$$\langle X, X \rangle_T = \int_0^T \sigma_t^2 dt \text{ and } \langle f(X), f(X) \rangle_T = \int_0^T f'(X_t)^2 \sigma_t^2 dt$$

(see Theorem 29 (page 75-76) of Protter (2004)). One interesting question arises immediately – which volatility are the volatility estimators estimating? When we make use of the observations Y_{t_i} to estimate the volatility, we might think that we are estimating $\langle X, X \rangle_T$ because Y_{t_i} is just the contaminated version of X_{t_i}; but given the X process, Y_{t_i} is centered at $f(X_{t_i})$, rather than X_{t_i}. We note that since both X_t and $f(X_t)$ are Itô processes, without further model assumptions, we have nothing in the model that can answer the question of which volatility is the true underlying one.

These two volatilities $\langle X, X \rangle_T$ and $\langle f(X), f(X) \rangle_T$ are often similar quantities if $f(x) \approx x$, which makes it not so crucial to think about the above question; but this may not always be the case. Our first objective is to make it clear which volatility the volatility estimators are estimating, and how good the approximations are.

TSRV is a typical example of volatility estimators. For the moment, we focus on determining the properties of TSRV. We shall make use of some of the notations from Zhang, et al. (2005):

Let $\mathcal{G} = \{0 = t_0, t_1, t_2, \cdots, t_n = T\}$ be the grid containing all the observation times. We suppose $\mathcal{G}$ is partitioned into K non-overlapping subgrids $\mathcal{G}^{(k)}, k = 1, \cdots, K$. As introduced in Zhang, et al. (2005), a typical example of selecting the subgrids is to use the regular allocation:

$$\mathcal{G}^{(k)} = \{t_{k-1}, t_{k-1+K}, t_{k-1+2K}, \cdots, t_{k-1+n_kK}\}.$$

Let $n_k = |\mathcal{G}^{(k)}|$, the integer making $t_{k-1+n_k K}$ the last element in $\mathcal{G}^{(k)}$; and $\bar{n} = \frac{1}{K}\sum_{k=1}^{K} n_k = \frac{1}{K}(n-K+1)$. Further define $[Z^{(1)}, Z^{(2)}]_T^{(all)} = [Z^{(1)}, Z^{(2)}]_T^{\mathcal{G}}$ and $[Z^{(1)}, Z^{(2)}]_T^{(avg)} = \frac{1}{K}\sum_{k=1}^{K}[Z^{(1)}, Z^{(2)}]_T^{\mathcal{G}^{(k)}}$ for two processes $Z^{(1)}$ and $Z^{(2)}$.

Estimators of the form $[Y,Y]_T^{\mathcal{H}}$ with $\mathcal{H} \subset \mathcal{G}$ are usually known as the RV. The TSRV is given by

$$\widehat{\langle X, X\rangle}_T = [Y,Y]_T^{(avg)} - \frac{\bar{n}}{n}[Y,Y]_T^{(all)}. \tag{6}$$

We shall assume constant step size ($\Delta t_i = T/n$) and that

$$\text{as } n \to \infty, K \to \infty \text{ and } n/K \to \infty. \tag{7}$$

Note that our results generalize quite predictably if we allow Δt_i to vary (see, the theory in Zhang, et al. (2005)).

2.2 Estimators of Volatilities - Estimators of $\langle f(X), f(X)\rangle_T$.

Denote

$$\epsilon_{t_i} = Y_{t_i} - f(X_{t_i}); \tag{8}$$

note that under (2), the conditional moments of ϵ_{t_i} only depend on the value of X_{t_i}. We assume the conditional 2nd moment of ϵ_{t_i} is continuous, and there exists $\delta_0 > 0$, such that the conditional $4 + 2\delta_0$'th moment of ϵ_{t_i} is bounded on compact sets; that is:

$$g(x) := E(\epsilon_{t_i}^2 | X_{t_i} = x) \text{ is continuous,} \tag{9}$$

$$\forall l > 0, \exists M_{(4+2\delta_0, l)}, s.t. E(|\epsilon_{t_i}|^{4+2\delta_0}|X_{t_i} = x) \leq M_{(4+2\delta_0, l)}, \text{ when } x \in [-l, l]. \tag{10}$$

We shall need the concept of stable convergence, as follows.

Definition 3. *Consider the σ-field $\Xi = \sigma(X_s, 0 \leq s \leq T)$. We say that a sequence ζ_n converges stably to ζ provided, for all $F \in \Xi$ and all bounded continuous g, $EI_F g(\zeta_n) \to EI_F g(\zeta)$ as $n \to \infty$, where ζ is defined on an extension of the original space.*

Note that since X is continuous, the stable convergence of a sequence ζ_n is equivalent to the joint convergence of ζ_n with the process $X_s, 0 \leq s \leq T$, see, Section 2 of Jacod & Protter (1998). This would not have been the case if X were discontinuous. Also, note that we are using a specific reference σ-field Ξ, which is a little different from standard usage.

Theorem 1. *When we take $K = cn^{2/3}$ (the best possible order of TSRV), under the assumptions (5), (9) and (10),*

$$n^{\frac{1}{6}}(\widehat{\langle X, X\rangle}_T - \langle f(X), f(X)\rangle_T) = n^{\frac{1}{6}}(\widehat{\langle X, X\rangle}_T - \int_0^T f'(X_t)^2 \sigma_t^2 dt)$$

$$\to_{\mathcal{L}} (\frac{8}{Tc^2}\int_0^T g(X_t)^2 dt + c\xi^2 T)^{1/2} N(0,1)$$

stably, where

$$\xi^2 = \frac{4}{3} \int_0^T (f'(X_t)\sigma_t)^4 dt. \quad \blacksquare \tag{11}$$

It is clear from this result what changes and what doesn't change for this more general contamination, compared to the case of independent additive error studied in Zhang, et al. (2005).

- The volatility which is being estimated is that of $f(X_t)$. (In Zhang, et al. (2005), $f(x) = x$.)

- The rate of convergence $n^{1/6}$ is the same as for independent additive error.

- The asymptotic variance changes to reflect the more complex form of contamination.

In summary, if we are happy to estimate the volatility of $f(X_t)$, the TSRV is exceedingly robust. The point about asymptotic variance is only an issue if one wishes to set an interval around the observation. As can be seen from Zhang, et al. (2005), this is difficult even with straight additive contamination.

2.3 Proof of Theorem 1

We need to do some preparations before proving Theorem 1.

First, note that under the assumption (10), the following is true:

$$\forall \theta < 4 + 2\delta_0, E(|\epsilon_{t_i}|^\theta|X_{t_i} = x) \text{ is bounded on } [-l, l]; \text{ (we write the bound as } M_{(\theta,l)}) \text{ and} \tag{12}$$

$$Var(\epsilon_{t_i}^2|X_{t_i} = x) = E(\epsilon_{t_i}^4|X_{t_i} = x) - E^2(\epsilon_{t_i}^2|X_{t_i} = x) \text{ is bounded on } [-l, l]; \text{ (say, by } M_{(Var,l)}). \tag{13}$$

And we shall use these notations in the proof:

$$M_T^{(1)} = \frac{1}{\sqrt{n}} \sum_{t_i \in \mathcal{G}} (\epsilon_{t_i}^2 - E(\epsilon_{t_i}^2|X)); \quad M_T^{(2)} = \frac{1}{\sqrt{n}} \sum_{t_i \in \mathcal{G}} \epsilon_{t_i} \epsilon_{t_{i-1}}; \quad M_T^{(3)} = \frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{t_i \in \mathcal{G}^{(k)}} \epsilon_{t_i} \epsilon_{t_{i,-}}$$

where $t_{i,-}$ denotes the previous element in $\mathcal{G}^{(k)}$ when $t_i \in \mathcal{G}^{(k)}$. $\epsilon_{t_{-1}} = 0$, $\epsilon_{t_{i,-}} = 0$ for $t_i = \min \mathcal{G}^{(k)}$.

Proposition 1. *Assume that $E(|A_n| |X)$ is $O_P(1)$, then A_n is $O_P(1)$.*

Proof.

$$\begin{aligned}
P(|A_n| > K) &\leq P(|A_n| I_{\{E(|A_n| |X) \leq K'\}} > K) + P(E(|A_n| |X) > K') \\
&\leq \frac{E(|A_n| I_{\{E(|A_n| |X) \leq K'\}})}{K} + P(E(|A_n| |X) > K') \\
&= \frac{E(E(|A_n| |X) I_{\{E(|A_n| |X) \leq K'\}})}{K} + P(E(|A_n| |X) > K') \\
&\leq \frac{K'}{K} + P(E(|A_n| |X) > K')
\end{aligned}$$

for all K, K', hence the result follows.

Lemma 1.

$$[Y, Y]_T^{(all)} = [\epsilon, \epsilon]_T^{(all)} + O_P(1); \tag{14}$$

$$[Y, Y]_T^{(avg)} = [\epsilon, \epsilon]_T^{(avg)} + [f(X), f(X)]_T^{(avg)} + O_P(\frac{1}{\sqrt{K}}). \tag{15}$$

Proof. Define $\tau_l = \inf\{t : |X_t| \geq l\}$, $\forall\, l$. Note that τ_l has the property that

$$P(\tau_l \leq T) \to 0 \text{ as } l \to \infty. \tag{16}$$

Also define $\Delta f(X_{t_i})) = f(X_{t_{i+1}}) - f(X_{t_i})$, for $i = 0, 1, \cdots, n-1$.

By (12), $g(X_t) = E(\epsilon_t^2|X_t)$, $t \leq T$ is bounded by $M_{(2,l)}$ on $\{\tau_l > T\}$.

$$E(([f(X), \epsilon]_T^{(all)})^2 I_{\{\tau_l > T\}}|X)$$

$$= I_{\{\tau_l > T\}} \sum_{i=1}^{n-1} (\Delta f(X_{t_{i-1}}) - \Delta f(X_{t_i}))^2 E(\epsilon_{t_i}^2|X) + \Delta f(X_{t_{n-1}})^2 E(\epsilon_{t_n}^2|X) + \Delta f(X_{t_0})^2 E(\epsilon_{t_0}^2|X)$$

$$\leq I_{\{\tau_l > T\}} M_{(2,l)} [\sum_{i=1}^{n-1} (\Delta f(X_{t_{i-1}}) - \Delta f(X_{t_i}))^2 + \Delta f(X_{t_{n-1}})^2 + \Delta f(X_{t_0})^2]$$

$$= 2 I_{\{\tau_l > T\}} M_{(2,l)} ([f(X), f(X)]_T^{(all)} - \sum_{i=1}^{n-1} \Delta f(X_{t_{i-1}}) \Delta f(X_{t_i}))$$

$$\leq 4 I_{\{\tau_l > T\}} M_{(2,l)} [f(X), f(X)]_T^{(all)}$$

$$= O_P(1),$$

where we have used the Cauchy-Schwarz Inequality. Hence by Proposition 1 and (16),

$$[f(X), \epsilon]_T^{(all)} = O_P(1). \tag{17}$$

Parallel argument shows that $E(([f(X), \epsilon]_T^{(avg)})^2 I_{\{\tau_l > T\}}|X) = O_p(\frac{1}{K})$. Hence

$$[f(X), \epsilon]_T^{(avg)} = O_P(\frac{1}{\sqrt{K}}). \tag{18}$$

(17) and (18) imply (14) and (15) because

$$[Y, Y]_T^{(all)} = [\epsilon, \epsilon]_T^{(all)} + [f(X), f(X)]_T^{(all)} + 2[f(X), \epsilon]_T^{(all)}$$

and

$$[Y, Y]_T^{(avg)} = [f(X), f(X)]_T^{(avg)} + [\epsilon, \epsilon]_T^{(avg)} + 2[f(X), \epsilon]_T^{(avg)}.$$

Lemma 2. $(M_T^{(2)}, M_T^{(3)})$ *are asymptotically independently normal conditionally on* X, *both with variance* $\frac{1}{T} \int_0^T g(X_t)^2 dt$.

Proof. We use $\langle \cdot, \cdot \rangle_T$ to denote the discrete-time predictable quadratic variations and covariations (see page 51 in Hall & Heyde (1980)) in this proof (note that they are different from the continuous time quadratic variations in Definition 2).

$(M_T^{(2)}, M_T^{(3)})$ are the end points of martingales with respect to filtration $\mathcal{F}_i = \sigma(\epsilon_{t_j}, j \leq i, X_t, \text{ all } t)$.

$$\langle M^{(2)}, M^{(2)} \rangle_T = \frac{1}{n} \sum_{t_i \in \mathcal{G}} Var(\epsilon_{t_i} \epsilon_{t_{i-1}} | \mathcal{F}_{i-1})$$

$$= \frac{1}{n} \sum_{t_i \in \mathcal{G}} \epsilon_{t_{i-1}}^2 g(X_{t_i})$$

$$= \frac{1}{n} \sum_{t_i \in \mathcal{G}} (\epsilon_{t_{i-1}}^2 - g(X_{t_{i-1}})) g(X_{t_i}) + \frac{1}{T} \sum_{t_i \in \mathcal{G}} g(X_{t_{i-1}}) g(X_{t_i}) \Delta t. \tag{19}$$

Note that

$$E((\frac{1}{n} \sum_{t_i \in \mathcal{G}} (\epsilon_{t_{i-1}}^2 - g(X_{t_{i-1}})) g(X_{t_i}) I_{\{\tau_l > T\}})^2 | X) = Var(\frac{1}{n} \sum_{t_i \in \mathcal{G}} (\epsilon_{t_{i-1}}^2 - g(X_{t_{i-1}})) g(X_{t_i}) I_{\{\tau_l > T\}} | X)$$

$$= \frac{1}{n^2} \sum_{t_i \in \mathcal{G}} Var(\epsilon_{t_{i-1}}^2 | X) g^2(X_{t_i}) I_{\{\tau_l > T\}}$$

$$\leq \frac{1}{n^2} \sum_{t_i \in \mathcal{G}} M_{(Var,l)} M_{(2,l)}$$

$$= O_P(1/n),$$

hence by Proposition 1 and (16), the first term of (19) $\frac{1}{n} \sum_{t_i \in \mathcal{G}} (\epsilon_{t_{i-1}}^2 - g(X_{t_{i-1}})) g(X_{t_i}) \rightarrow_P 0$. Therefore, by (9),

$$\langle M^{(2)}, M^{(2)} \rangle_T = \frac{1}{T} \sum_{t_i \in \mathcal{G}} g(X_{t_{i-1}}) g(X_{t_i}) \Delta t + o_p(1) \rightarrow_P \frac{1}{T} \int_0^T g(X_t)^2 dt.$$

Parallel argument shows

$$\langle M^{(3)}, M^{(3)} \rangle_T = \frac{1}{n} \sum_{k=1}^K \sum_{t_i \in \mathcal{G}^{(k)}} Var(\epsilon_{t_i} \epsilon_{t_{i,-}} | \mathcal{F}_{i-1}) \rightarrow_P \frac{1}{T} \int_0^T g(X_t)^2 dt.$$

On the other hand,

$$\langle M^{(2)}, M^{(3)} \rangle_T = \frac{1}{n} \sum_{k=1}^K \sum_{t_i \in \mathcal{G}^{(k)}} Cov(\epsilon_{t_i} \epsilon_{t_{i-1}}, \epsilon_{t_i} \epsilon_{t_{i,-}} | \mathcal{F}_{i-1}) = \frac{1}{n} \sum_{k=1}^K \sum_{t_i \in \mathcal{G}^{(k)}} \epsilon_{t_{i-1}} \epsilon_{t_{i,-}} E(\epsilon_{t_i}^2 | X).$$

As a consequence,

$$
\begin{aligned}
E((\langle M^{(2)}, M^{(3)}\rangle_T)^2 I_{\{\tau_l > T\}}|X) =& \frac{1}{n^2} \sum_{k=1}^{K} \sum_{t_i \in \mathcal{G}^{(k)}} E^2(\epsilon_{t_i}^2|X) E(\epsilon_{t_{i-1}}^2 \epsilon_{t_{i,-}}^2 I_{\{\tau_l > T\}}|X) \\
\leq& \frac{1}{n^2} \sum_{k=1}^{K} \sum_{t_i \in \mathcal{G}^{(k)}} E^2(\epsilon_{t_i}^2|X) \sqrt{E(\epsilon_{t_{i-1}}^4 I_{\{\tau_l > T\}}|X) E(\epsilon_{t_{i,-}}^4 I_{\{\tau_l > T\}}|X)} \\
\leq& \frac{1}{n} M_{(2,l)}^2 M_{(4,l)} \\
=& O_P(\frac{1}{n}).
\end{aligned}
$$

By Proposition 1 and (16), $\langle M^{(2)}, M^{(3)}\rangle_T \to_P 0$.

By assumption (10) and Proposition 1, one can easily see that the conditional Lyapounov's conditions are satisfied. Also note that the limiting conditional variance $\frac{1}{T} \int_0^T g(X_t)^2 dt$ and the limiting conditional covariance 0 are measurable in (the completions of) all the $\sigma-$fields $\mathcal{F}_i$, one can make use of the Remarks right after the Corollary 3.1 (the martingale central limit theorem) in Hall & Heyde (1980) to obtain the conclusion.

Proof of Theorem 1.

Denote

$$
R_1 = (\epsilon_{t_0}^2 - E(\epsilon_{t_0}^2|X)) + (\epsilon_{t_n}^2 - E(\epsilon_{t_n}^2|X))
$$

and

$$
R_2 = \sum_{k=1}^{K} [(\epsilon_{min\mathcal{G}^{(k)}}^2 - E(\epsilon_{min\mathcal{G}^{(k)}}^2|X)) + (\epsilon_{max\mathcal{G}^{(k)}}^2 - E(\epsilon_{max\mathcal{G}^{(k)}}^2|X))].
$$

By (13), $E(R_1^2 I_{\{\tau_l > T\}}|X) = O_p(1)$. Hence $R_1 = O_p(1)$ by Proposition 1 and (16). Similarly, $E(R_2^2 I_{\tau_l > T}|X) = O_p(K)$, hence $R_2 = O_p(K^{1/2})$. As a consequence,

$$
\begin{aligned}
[\epsilon, \epsilon]_T^{(all)} =& 2 \sum_{t_i \in \mathcal{G}} (\epsilon_{t_i}^2 - E(\epsilon_{t_i}^2|X)) - 2 \sum_{t_i > 0} \epsilon_{t_i} \epsilon_{t_{i-1}} + 2 \sum_{t_i \in \mathcal{G}} E(\epsilon_{t_i}^2|X) - R_1 \\
=& 2\sqrt{n}(M^{(1)} - M^{(2)}) + 2 \sum_{t_i \in \mathcal{G}} E(\epsilon_{t_i}^2|X) + O_P(1)
\end{aligned}
$$

and

$$
\begin{aligned}
K[\epsilon, \epsilon]_T^{(avg)} =& 2\sqrt{n}(M^{(1)} - M^{(3)}) - R_2 + 2 \sum_{k=1}^{K} \sum_{t_i \in \mathcal{G}^{(k)}} E(\epsilon_{t_i}^2|X) \\
=& 2\sqrt{n}(M^{(1)} - M^{(3)}) + O_P(K^{1/2}) + 2 \sum_{k=1}^{K} \sum_{t_i \in \mathcal{G}^{(k)}} E(\epsilon_{t_i}^2|X).
\end{aligned}
$$

Therefore, conditionally on the X process,

$$
\begin{aligned}
\frac{K}{\sqrt{n}}([\epsilon,\epsilon]_T^{(avg)} - \frac{\bar{n}}{n}[\epsilon,\epsilon]_T^{(all)}) &\approx \frac{1}{\sqrt{n}}(K[\epsilon,\epsilon]_T^{(avg)} - [\epsilon,\epsilon]_T^{(all)}) \\
&= (2(M^{(2)} - M^{(3)})) + O_P(\sqrt{\frac{K}{n}}) \\
&\to_{\mathcal{L}} N(0, \frac{8}{T}\int_0^T g(X_t)^2 dt).
\end{aligned}
\tag{20}
$$

Observe that

$$
\begin{aligned}
\widehat{\langle X, X \rangle}_T &= [Y,Y]_T^{(avg)} - \frac{\bar{n}}{n}[Y,Y]_T^{(all)} \\
\text{(by Lemma 1)} &= [f(X),f(X)]_T^{(avg)} + [\epsilon,\epsilon]_T^{(avg)} + O_p(\frac{1}{\sqrt{K}}) - \frac{\bar{n}}{n}[\epsilon,\epsilon]_T^{(all)} - O_p(\frac{\bar{n}}{n}) \\
&= [f(X),f(X)]_T^{(avg)} + [\epsilon,\epsilon]_T^{(avg)} - \frac{\bar{n}}{n}[\epsilon,\epsilon]_T^{(all)} + O_p(\frac{1}{\sqrt{K}})
\end{aligned}
$$

and note that $\frac{K}{\sqrt{n}} \sim \frac{K}{\sqrt{K\bar{n}}} \sim \sqrt{\frac{K}{\bar{n}}}$, by (20), conditionally on the X process,

$$
\sqrt{\frac{K}{\bar{n}}}(\widehat{\langle X, X \rangle}_T - [f(X),f(X)]_T^{(avg)}) \to_{\mathcal{L}} N(0, \frac{8}{T}\int_0^T g(X_t)^2 dt).
\tag{21}
$$

On the other hand, $f(X_t)$ is a semimartingale, $df(X_t) = (f'(X_t)\mu_t + \frac{1}{2}f''(X_t)\sigma_t^2)dt + f'(X_t)\sigma_t dB_t$. By Zhang, et al. (2005),

$$
\sqrt{\frac{n}{K}}([f(X),f(X)]_T^{(avg)} - \langle f(X),f(X) \rangle_T) \to_{\mathcal{L}} \xi\sqrt{T} \cdot Z_{discrete},
\tag{22}
$$

where ξ is defined as in (11), and $Z_{discrete} \sim N(0,1)$, independent of the process X. The convergence in law is stable.

Combining $(21), (22)$, one has

$$
\begin{aligned}
\widehat{\langle X, X \rangle}_T - \langle f(X),f(X) \rangle_T &= (\widehat{\langle X, X \rangle}_T - [f(X),f(X)]_T^{(avg)}) + ([f(X),f(X)]_T^{(avg)} - \langle f(X),f(X) \rangle_T) \\
&= O_p(\frac{\bar{n}^{1/2}}{K^{1/2}}) + O_p(\bar{n}^{-1/2}).
\end{aligned}
$$

The error is minimized when $K = O(n^{2/3})$. If we take $K = cn^{2/3}$, we have (by exploiting the conditional convergence in (21))

$$
n^{\frac{1}{6}}(\widehat{\langle X, X \rangle}_T - \langle f(X),f(X) \rangle_T) \to_{\mathcal{L}} (\frac{8}{Tc^2}\int_0^T g(X_t)^2 dt + c\xi^2 T)^{1/2}N(0,1) \text{ stably.}
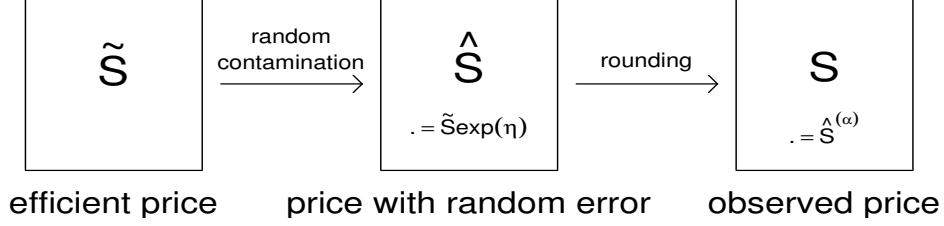$$

Figure 1: Two stage contamination: random error followed by rounding

3 A CASE STUDY

3.1 Another Form of Contamination

Two types of errors – additive errors and rounding errors have been proposed to be candidates of market microstructure errors. Results of when one of them plays the role are available. Now, we consider the case when both types of errors are present.

Suppose at the transaction times, the latent return process X_{t_i} is contaminated by an independent additive error process η_{t_i}, and then rounded to reflect that prices are quotes on a grid (typically in multiples of one cent). We thus envisage a two stage procedure where a latent efficient price $\tilde{S} = \exp(X)$ is first subjected to multiplicative random error: $\hat{S} = \tilde{S}\exp(\eta)$. The actual price S is then the rounded value of $\hat{S}$. If we take, as usual, $Y = \log S$, our final product is the observed process Y_{t_i} of rounded contaminated prices:

$$Y_{t_i} = \log((\exp(X_{t_i} + \eta_{t_i}))^{(\alpha)}),$$

where $s^{(\alpha)} = \alpha[s/\alpha]$ is the value of s rounded to the nearest multiple of α. The model is somewhat similar to that used by Large (2005). It can be illustrated as in Figure 1.

For practical purposes, we further assume that the smallest observation of the security price is α, which makes the observations of the log prices have the form

$$Y_{t_i} = \log \alpha \vee \log((\exp(X_{t_i} + \eta_{t_i}))^{(\alpha)}). \tag{23}$$

We consider the case when the random errors are independent identically distributed normal random variables, with mean 0 and positive variance; that is

$$\eta_{t_i} \sim_{i.i.d.} N(0, \gamma^2), \quad \gamma > 0. \tag{24}$$

In this case,

$$f(x) = E(Y_{t_i}|X_{t_i} = x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\gamma}(\log \alpha \vee \log((e^z)^{(\alpha)}))e^{-\frac{(z-x)^2}{2\gamma^2}}\,dz \tag{25}$$

is a twice continuously differentiable function. And the assumptions (9) and (10) hold. Therefore, by theorem 1, the TSRV is a robust estimator of $\langle f(X), f(X)\rangle_T$.

In studying this, we assume that α is a fixed quantity independent of the number of observations. Note that in the case where $\alpha \to 0$, one can expect relatively well posed behavior, in view of Kolassa & McCullagh (1990) and Delattre & Jacod (1997).

3.2 Robustness Works: When γ is Big:

Assume that the latent price process $\tilde{S} = \exp(X_t)$ has small probability to go below α for $t \in [0, T]$, under model (23) and assumption (24), one has,

$$f(X_t) \approx X_t, \text{ and } f'(X_t) \approx 1 \text{ for } t \in [0, T], \quad \text{for suitably big } \gamma\text{'s.}$$

By "suitably big γ's", we mean the size of the random error is big enough so that the possibility that it pull the observations of the prices up or down several grid points (multiples of α) is not negligible. In this case, when taking the conditional expectation, the positive and negative errors cancel out. And this leads to the result that $f(X_t) \approx X_t$.

In this case, $\langle f(X), f(X)\rangle_T \approx \langle X, X\rangle_T$. Therefore, the TSRV, which is a robust estimator of $\langle f(X), f(X)\rangle_T$, is a good estimator of $\langle X, X\rangle_T$ as well.

These relationships are illustrated in section 3.4.

3.3 How Things Can Go Wrong: When $\gamma \to 0$:

When γ is small but not 0, by Theorem 1, we know that the TSRV goes to the limit $\langle f(X), f(X)\rangle_T$ robustly. But this volatility $\langle f(X), f(X)\rangle_T$ is no longer close to $\langle X, X\rangle_T$.

To study the limiting behavior of $\langle f(X), f(X)\rangle_T$, we relate it to the local time L_T^a of the semimartingale X (for definition of the local time, see page 222, Revuz & Yor (1999)). By Revuz & Yor (1999) page 224 Corollary 1.6, for any positive Borel function Φ,

$$\int_0^t \Phi(X_s)d\langle X, X\rangle_s = \int_{-\infty}^{\infty} \Phi(a)L_t^a\,da. \tag{26}$$

And by Corollary 1.8 page 226 in the same book, the family L^a may be chosen such that

$$a \to L_t^a \text{ is Hölder continuous of order } \beta, \text{ for every } \beta < 1/2, \tag{27}$$

and uniformly in t on every compact interval. We shall consider only the version of the local time L_t^a which satisfies the condition (27).

Relating $\langle f(X), f(X)\rangle_T$ to L_T^a by (26), one has the following result:

Theorem 2. *As $\gamma \to 0$, almost surely,*

$$\gamma \langle f(X), f(X) \rangle_T \to \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)} (\log \frac{k+1}{k})^2,$$

where L_t^a is the local time of the continuous semimartingale X. ∎

In other words, the "target" $\langle f(X), f(X) \rangle_T$ which we are estimating blows up as γ goes to zero, and is of order $1/\gamma$. This raises questions of whether $\langle f(X), f(X) \rangle_T$ is, in this case, the quantity that we are really after.

3.4 Illustration

We take a typical sample path to illustrate the situation: suppose the latent log price process X_t follows (1) with $\mu_t = 0$ and $\sigma_t = 0.2$, $\forall t \in [0, \infty)$ (these are the annualized parameters). Suppose at the observation time t_i, the price $\exp(X_{t_i})$ is first contaminated by an independent multiplicative random error $\exp(\eta_{t_i})$ with η_{t_i} independent identically distributed as $N(0, \gamma^2)$, and then rounded to the nearest multiple of $\alpha = 0.01$ (one cent). The quantity of interest is the volatility of the process over time $t \in [0, T]$ with $T = 1/252$ (i.e., one day). We assume that a day consists of 6.5 hours of open trading, and the price process is observed once every second (n=23400).

A sample path of the latent log price process X_t, $t \in [0, T]$ is plotted in Figure 2, together with its corresponding pure rounded process (the solid line) and two $f(X_t)$ processes (see (25)) with $\gamma = 0.001$ ("*"'s) and $\gamma = 0.005$ ("o"'s) respectively.

Figure 3 records the TSRV of this particular sample path X_t, $t \in [0, T]$ in Figure 2, with random contaminations of different sizes (with standard error γ ranges from 0.0002 to 0.006). The solid line is the volatility $\langle X, X \rangle_T$.

One sees from Figure 2 that for this process, when γ is as big as 0.005, the $f(X)$ is close to X. While when γ is smaller, the process $f(X)$ diverges from X; in fact, it goes closer to the (discontinuous) pure rounded process. Figure 3 shows that when γ is suitably big, the TSRV can be a good estimator of $\langle X, X \rangle_T$, but when γ is too small, the estimator is not estimating $\langle X, X \rangle_T$, but rather a much larger quantity. Note that though similarly shaped, this graph is not a signature plot in the sense of Andersen, et al. (2000), since the horizontal axis represents γ rather than sample size. There is, however, a connection between these two types of plots, as shown in equation (29) below.

3.5 How Error and Sample Size Relate to Each Other – Comparison to Case When $\gamma = 0$:

When $\gamma = 0$, the additive error is gone, only the rounding error is present. In this case, the observations are themselves the conditional expectations, and $f(x)$ is no longer continuous.

$$Y_{t_i} = f(X_{t_i}) = E(Y_{t_i} | X_{t_i}) = \log \alpha \vee \log((exp(X_{t_i}))^{(\alpha)}). \tag{28}$$
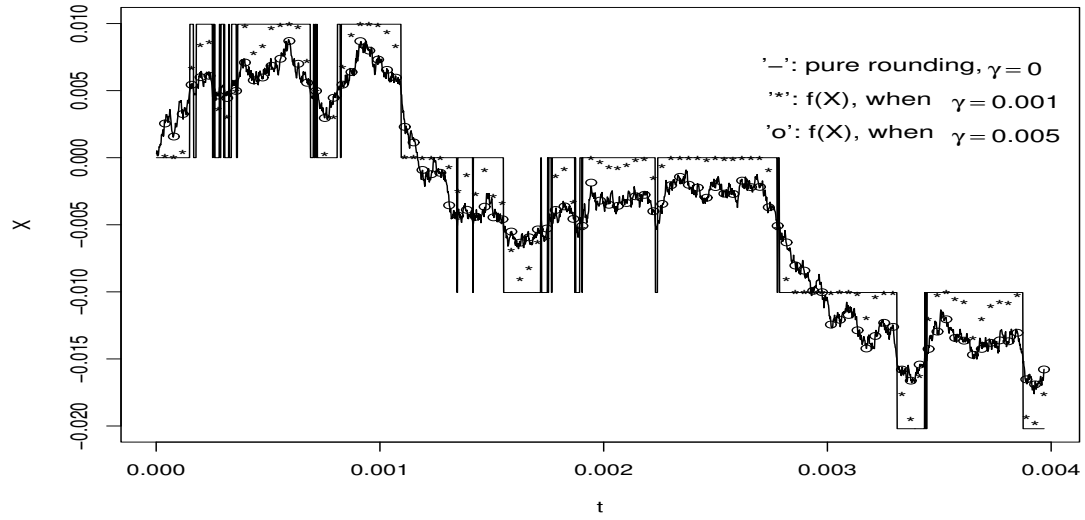
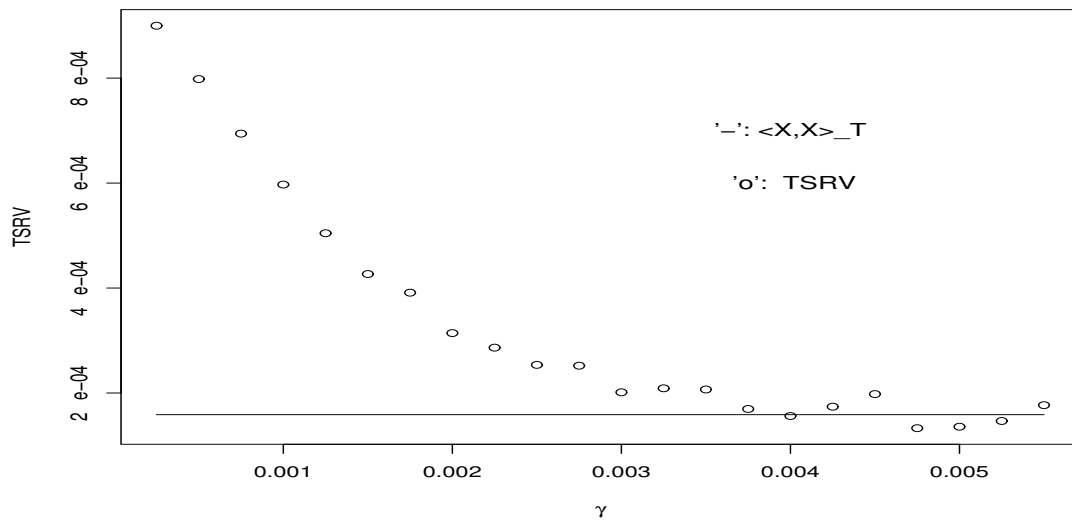Figure 2: Relationship between $f(X)$ and X on one (random) sample path.

Figure 3: TSRV *vs* size of the random contamination, based on one (random) latent log price process ($\langle X, X \rangle_T = (0.2)^2/252 \approx 1.59 * 10^{-4}$). For more details of the simulation, please refer to Section 3.4.

A modification of Jacod (1996)'s proof gives the following result: (Recall the notations from section 2.2 about the TSRV $\langle \widehat{X, X} \rangle_T$. In particular, K is the number of subgrids and $\bar{n} = \frac{1}{K}(n - K + 1)$ is the average number of elements in the subgrids. Also recall assumption (7), which is equivalent to "$\bar{n} \to \infty$ and $\bar{n}/n \to 0$ as $n \to \infty$".)

Theorem 3. *When $X = \sigma W$, where W is a standard Brownian Motion, one has*

$$plim_{n\to\infty} \frac{1}{\sqrt{\bar{n}}}\langle \widehat{X, X} \rangle_T = \frac{1}{\sigma\sqrt{T}}\sqrt{\frac{2}{\pi}}\sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)}(\log\frac{k+1}{k})^2. \quad \blacksquare$$

We can see from Theorems 2 and 3 that to first order,

TSRV under pure rounding and no contamination

$$= \sqrt{\frac{8\bar{n}\gamma^2}{\sigma^2 T}} \times \text{TSRV under rounding after contamination of size } \gamma \qquad (29)$$

Thus, in a sense, contamination plays a rôle slightly similar to sample size when there is no contamination:

$$\gamma^{-2} \text{ under random contamination } \approx \bar{n}\frac{8}{\sigma^2 T} \text{ under no random contamination.} \qquad (30)$$

In both cases, the size of of γ^{-2} and $\bar{n}$ have similar functions in quantifying the ill-posedness of the respective estimation problems. The deeper meaning of this remains, for the moment, a little mysterious even to the authors.

3.6 Proofs of Theorem 2 and Theorem 3

Proof of Theorem 2.

One has, by (26), for $\mu(da) = (f'(a))^2 da$,

$$\langle f(X), f(X) \rangle_T = \int_0^T (f'(X_t))^2 d\langle X, X \rangle_t = \int_{-\infty}^{\infty} L_T^a \mu(da).$$

Recall that $f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\gamma}(\log\alpha \vee \log((e^z)^{(\alpha)}))e^{-\frac{(z-x)^2}{2\gamma^2}}dz$, hence,

$$f'(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\gamma}(\log\alpha \vee \log((e^z)^{(\alpha)}))\frac{z-x}{\gamma^2}e^{-\frac{(z-x)^2}{2\gamma^2}}dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\gamma}\log\alpha \vee \log(\exp(x+\gamma v)^{(\alpha)})\frac{v}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}dv.$$

For $k = 1, 2, 3, \cdots$, one has, $\forall y \in \mathbb{R}$,

$$\gamma f'(\log((k+\tfrac{1}{2})\alpha) + y\gamma) = \int_{-\infty}^{\infty} \log \alpha \vee \log(\exp(\log((k+\tfrac{1}{2})\alpha) + y\gamma + v\gamma)^{(\alpha)}) \frac{v}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv$$

$$= E_V\left(\log \alpha \vee \log(((k+\tfrac{1}{2})\alpha \cdot e^{\gamma(y+V)})^{(\alpha)}) \cdot V\right), \quad V \sim N(0,1).$$

By the Dominated Convergence Theorem,

$$\lim_{\gamma \to 0} \gamma f'(\log((k+\tfrac{1}{2})\alpha) + y\gamma) = E_V\left(\log((k+1)\alpha) \cdot VI\{y+V>0\}\right) + E_V\left(\log(k\alpha) \cdot VI\{y+V<0\}\right)$$

$$= \log((k+1)\alpha) \frac{1}{\sqrt{2\pi}} \int_{-y}^{\infty} z e^{-\frac{z^2}{2}} dz + \log(k\alpha) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-y} z e^{-\frac{z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} (\log((k+1)\alpha) - \log(k\alpha))$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \log(\frac{k+1}{k}). \tag{31}$$

A similar argument shows that

$$\forall y \in \mathbb{R}, \quad \gamma f'(\log \alpha + y\gamma) \to 0 \text{ as } \gamma \to 0. \tag{32}$$

By (31), for $k = 1, 2, 3 \cdots$,

$$\lim_{\gamma \to 0} \int_{\log((k+\frac{1}{2})\alpha)-n\gamma}^{\log((k+\frac{1}{2})\alpha)+n\gamma} \gamma(f'(x))^2 dx = \lim_{\gamma \to 0} \int_{-n}^{n} (\gamma f'(\log((k+\tfrac{1}{2})\alpha) + y\gamma))^2 dy$$

$$= \int_{-n}^{n} (\frac{1}{\sqrt{2\pi}} (\log((\frac{k+1}{k})))^2 e^{-y^2} dy$$

$$= \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2 \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{2}n}^{\sqrt{2}n} e^{-\frac{z^2}{2}} dz.$$

For any $a \in [\log \alpha, \infty)$, $b \in (a, \infty]$, suppose the set $\{k : \log((k+\tfrac{1}{2})\alpha) \in (a, b)\}$ is not empty. Denote $k_a^0 = \lceil \frac{e^a}{\alpha} - \frac{1}{2} \rceil$, the smallest integer k such that $\log((k+\tfrac{1}{2})\alpha) > a$; $k_b^1 = \lfloor \frac{e^b}{\alpha} - \frac{1}{2} \rfloor$, the biggest integer k such that $\log((k+\tfrac{1}{2})\alpha) < b$.

Note that for any $N \le \min(\frac{\log((k_b^1+\frac{1}{2})\alpha) - \log((k_b^1-\frac{1}{2})\alpha)}{2\gamma}, \frac{\log((k_a^0+\frac{1}{2})\alpha) - a}{\gamma}, \frac{b - \log((k_b^1+\frac{1}{2})\alpha)}{\gamma})$,

$$\gamma \mu(a, b) \ge \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \gamma \mu\left(\log((k+\tfrac{1}{2})\alpha) - N\gamma, \log((k+\tfrac{1}{2})\alpha) + N\gamma\right)$$

$$= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{\log((k+\frac{1}{2})\alpha)-N\gamma}^{\log((k+\frac{1}{2})\alpha)+N\gamma} \gamma(f'(x))^2 dx.$$

Hence,

$$\liminf_{\gamma \to 0} \gamma \mu(a, b) \ge \lim_{\gamma \to 0} \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{\log((k+\frac{1}{2})\alpha)-N\gamma}^{\log((k+\frac{1}{2})\alpha)+N\gamma} \gamma(f'(x))^2 dx$$

$$= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2 (\Phi(\sqrt{2}N) - \Phi(-\sqrt{2}N)),$$

where Φ is the distribution function of a standard normal distribution.

As $\gamma \to 0$, N can be chosen to be arbitrarily large, therefore,

$$
\begin{aligned}
\liminf_{\gamma \to 0} \gamma \mu(a,b) &\geq \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2 (\Phi(\infty) - \Phi(-\infty)) \\
&= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2.
\end{aligned}
\tag{33}
$$

On the other hand, for any $\gamma > 0$, there exists M, such that

$$
\begin{aligned}
\gamma \mu(a,b) &\leq \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \gamma \mu \Big(\log((k+\frac{1}{2})\alpha) - M\gamma, \log((k+\frac{1}{2})\alpha) + M\gamma \Big) \\
&= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{\log((k+\frac{1}{2})\alpha)-M\gamma}^{\log((k+\frac{1}{2})\alpha)+M\gamma} \gamma(f'(x))^2 dx \\
&= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{-M}^{M} (\gamma f'(\log((k+\frac{1}{2})\alpha) + y\gamma))^2 dy \\
&\leq \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{-\infty}^{\infty} (\gamma f'(\log((k+\frac{1}{2})\alpha) + y\gamma))^2 dy,
\end{aligned}
$$

which implies,

$$
\begin{aligned}
\limsup_{\gamma \to 0} \gamma \mu(a,b) &\leq \lim_{\gamma \to 0} \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \int_{-\infty}^{\infty} (\gamma f'(\log((k+\frac{1}{2})\alpha) + y\gamma))^2 dy \\
&= \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2.
\end{aligned}
\tag{34}
$$

By (33) and (34), for all $(a,b) \subset (\log \alpha, \infty)$ such that $\{k : \log((k+\frac{1}{2})\alpha) \in (a,b)\} \neq \emptyset$,

$$
\lim_{\gamma \to 0} \gamma \mu(a,b) = \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2.
\tag{35}
$$

By linearity of measures, (35) implies that

$$
\lim_{\gamma \to 0} \gamma \mu(a,b) = 0 \text{ for } (a,b) \subset (\log \alpha, \infty) \text{ and } \{k : \log((k+\frac{1}{2})\alpha) \in (a,b)\} = \emptyset.
$$

By (32) and a similar argument as above, one has, $\lim_{\gamma \to 0} \gamma \mu(a,b) = 0$ for any $(a,b) \subset (-\infty, \log(\frac{3}{2}\alpha))$.

In summary,

$$
\lim_{\gamma \to 0} \gamma \mu(a,b) = \sum_{k:\log((k+\frac{1}{2})\alpha)\in(a,b)} \frac{1}{2\sqrt{\pi}} (\log \frac{k+1}{k})^2, \text{ for any } (a,b) \subset \mathbb{R}.
$$

Therefore, as $\gamma \to 0$, $\gamma\mu$, as a measure on $\mathbb{R}$, converges to a finite measure ν, which has point mass $\frac{1}{2\sqrt{\pi}}(\log\frac{k+1}{k})^2$ on $\log((k+\frac{1}{2})\alpha)$, $k = 1, 2, 3 \cdots$.

On the other hand, by (26),

$$\int_{-\infty}^{\infty} L_T^a da = \int_0^T d\langle X, X\rangle_t < \infty. \tag{36}$$

The integrability (36), together with the Hölder continuity (27) of L_T^a gives,

$$L_T^a \text{ is almost surely bounded in } a. \tag{37}$$

As a consequence, almost surely,

$$\lim_{\gamma\to 0} \gamma\langle f(X), f(X)\rangle_T = \lim_{\gamma\to 0} \gamma \int_{-\infty}^{\infty} L_T^a \mu(da) = \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)} (\log\frac{k+1}{k})^2.$$

Proof of Theorem 3.:

We borrow the notations from Jacod (1996): For $t_i = \frac{iT}{n}, i = 0, 1, \cdots n$,

$\xi_i^n := X_{t_i} - X_{t_{i-1}}$; $\chi_i^n := (f(X_{t_i}) - f(X_{t_{i-1}}))^2$, where $f(x)$ is defined in (28);
$R_n^k := \{(x,y) : f(x) = \log(k\alpha), f(y) = \log((k+1)\alpha) \text{ or } f(x) = \log((k+1)\alpha), f(y) = \log(k\alpha)\}$;
$R_n := \cup_{k=1}^{\infty} R_n^k$; $S_n := \mathbb{R}^2 \setminus R_n$; $T(a) := \{(x,y) : x < a \le y \text{ or } y < a \le x\}$;
$T_n := \cup_{k=1}^{\infty} T(\log((k+\frac{1}{2})\alpha))$; $\hat{W}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{R_n}(X_{t_{i-1}}, X_{t_i})$; $W_n := \sum_{i=1}^n \frac{1}{\sqrt{n}} \chi_i^n I_{S_n}(X_{t_{i-1}}, X_{t_i})$.

If $(x, y) \in S_n$, then either $f(x) = f(y)$,

$$\text{or} \quad |\exp(f(x)) - \exp(f(y))| > \alpha. \tag{38}$$

In the case of (38), without lost of generality, one can assume $\exp(x) > \exp(y) \ge \alpha$, then,

$$|f(x) - f(y)| = \log\frac{(\exp(x))^{(\alpha)}}{(\exp(y))^{(\alpha)}} \le \log\frac{\exp(x) + \alpha/2}{\exp(y) - \alpha/2} \le \log\frac{\exp(x) + \exp(x)/2}{\exp(y) - \exp(y)/2} = \log 3 + |x - y|. \tag{39}$$

If in addition we have that both $\exp(x)$ and $\exp(y)$ are bounded by $M > 0$, then, there exists θ, $(x \wedge y) \le \theta \le (x \vee y)$, such that $\alpha \le |\exp(x) - \exp(y)| = \exp(\theta)|x - y| \le M|x - y|$. This implies that $|x - y| \ge \frac{\alpha}{M}$, hence, $\log 3 \le \frac{M\log 3}{\alpha}|x - y|$. By (39),

$$|f(x) - f(y)| \le (\frac{M\log 3}{\alpha} + 1)|x - y|. \tag{40}$$

It is easy to see that (40) holds for all $(x, y) \in S_n$ such that $\exp(x), \exp(y)$ are bounded by M. Therefore,

$$E(W_n I_{\{\tau_{logM} > T\}}) = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} E(\chi_i^n I_{S_n}(X_{t_{i-1}}, X_{t_i}) I_{\{\tau_{logM} > T\}}) \leq \sum_{i=1}^{n} \frac{1}{\sqrt{n}} (\frac{M \log 3}{\alpha} + 1)^2 \frac{\sigma^2 T}{n},$$

which implies (also make use of (16))

$$W_n \to_P 0 \quad \text{as } n \to \infty. \tag{41}$$

On the other hand, for $i = 1, 2, \cdots n$,

$$E(I_{\{|\exp(X_{t_i}) - \exp(X_{t_{i-1}})| \geq \alpha\}})$$
$$\leq E(\exp(X_{\frac{iT}{n}}) - \exp(X_{\frac{(i-1)T}{n}}))^2 / \alpha^2$$
$$= (E \exp(2\sigma W_{\frac{iT}{n}}) + E \exp(2\sigma W_{\frac{(i-1)T}{n}}) - 2E \exp(2\sigma W_{\frac{(i-1)T}{n}}) E \exp(\sigma W_{\frac{T}{n}})) / \alpha^2$$
$$= \frac{1}{\alpha^2} \exp(\frac{2\sigma^2(i-1)T}{n})(\exp(\frac{2\sigma^2 T}{n}) + 1 - 2\exp(\frac{\sigma^2 T}{2n}))$$
$$\leq \frac{1}{\alpha^2} \exp(2\sigma^2 T) \sum_{k=1}^{\infty} ((2\sigma^2 T)^k - 2(\sigma^2 T/2)^k) / (k! n^k)$$
$$\leq \frac{1}{n} \left(\frac{\exp(2\sigma^2 T)(exp(2\sigma^2 T) - exp(\sigma^2 T/2))}{\alpha^2} \right),$$

which implies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\{|\exp(X_{t_i}) - \exp(X_{t_{i-1}})| \geq \alpha\}} \to_P 0 \quad \text{as } n \to \infty. \tag{42}$$

Note also that if $(X_{t_{i-1}}, X_{t_i}) \in R_n^k$, then $\chi_i^n = (\log \frac{k+1}{k})^2$. One has, for $k_l^1 = \lfloor \frac{e^l}{\alpha} - \frac{1}{2} \rfloor$ (the biggest integer k such that $\log((k + \frac{1}{2})\alpha) < l$),

$$\frac{1}{\sqrt{n}}[Y, Y]^{(all)} I_{\{\tau_l > T\}} = (\hat{W}_n \sum_{k=1}^{\infty} I_{R_n^k}(X_{t_{i-1}}, X_{t_i}) \chi_i^n + W_n) I_{\{\tau_l > T\}}$$

$$= (\sum_{k=1}^{k_l^1} (\hat{W}_n I_{R_n^k}(X_{t_{i-1}}, X_{t_i})) \cdot (\log \frac{k+1}{k})^2 + W_n) I_{\{\tau_l > T\}}. \tag{43}$$

One has $R_n \subset T_n$ and $|\exp(x) - \exp(y)| \geq \alpha$ when $(x, y) \in T_n \setminus R_n$. By (41), (42) and (43), one

sees that, for $\hat{W}'_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{T_n}(X_{t_{i-1}}, X_{t_i})$,

$$plim_{n\to\infty} \frac{1}{\sqrt{n}}[Y,Y]^{(all)} I_{\{\tau_l > T\}} = plim_{n\to\infty} \sum_{k=1}^{k_l^1} \hat{W}'_n I_{T(\log((k+\frac{1}{2})\alpha))}(X_{t_{i-1}}, X_{t_i})(\log\frac{k+1}{k})^2 I_{\{\tau_l > T\}}$$

$$= plim_{n\to\infty} \sum_{k=1}^{k_l^1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{T(\log((k+\frac{1}{2})\alpha))}(X_{t_{i-1}}, X_{t_i})(\log\frac{k+1}{k})^2 I_{\{\tau_l > T\}}$$

$$= \sum_{k=1}^{k_l^1} [(\log\frac{k+1}{k})^2 plim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{T(\log((k+\frac{1}{2})\alpha))}(X_{t_{i-1}}, X_{t_i})] I_{\{\tau_l > T\}}.$$

And for $k = 1, 2, 3, \cdots$, by Jacod (1996),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{T(\log((k+\frac{1}{2})\alpha))}(X_{t_{i-1}}, X_{t_i}) \to_{\mathbb{L}^2} \frac{1}{\sigma\sqrt{T}} \sqrt{\frac{2}{\pi}} L_T^{\log((k+\frac{1}{2})\alpha)}.$$

Hence, by letting l go to ∞, one has,

$$plim_{n\to\infty} \frac{1}{\sqrt{n}}[Y,Y]^{(all)} = \frac{1}{\sigma\sqrt{T}} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)}(\log\frac{k+1}{k})^2, \qquad (44)$$

and

$$plim_{\bar{n}\to\infty} \frac{1}{\sqrt{\bar{n}}}[Y,Y]^{(avg)} = \frac{1}{\sigma\sqrt{T}} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)}(\log\frac{k+1}{k})^2.$$

Applying these results to the TSRV (6), note that (44) implies $\frac{\sqrt{\bar{n}}}{n}[Y,Y]^{(all)} \to_P 0$ by assumption (7), one has

$$\frac{1}{\sqrt{\bar{n}}}\widehat{\langle X, X\rangle}_T = \frac{1}{\sqrt{\bar{n}}}([Y,Y]_T^{(avg)} - \frac{\bar{n}}{n}[Y,Y]_T^{(all)}) \to_P \frac{1}{\sigma\sqrt{T}} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{\infty} L_T^{\log((k+\frac{1}{2})\alpha)}(\log\frac{k+1}{k})^2.$$

4 CONCLUSION

We have shown in this paper that the robustness of the two scales realized volatility (TSRV) depends crucially on the deterministic part of the distortion through the function f defined in (4). On the other hand, in terms of consistency and order of convergence, the TSRV is always robust to the random part of the error $(Y - f(X))$. In Section 3, we have studied a particular model of contamination, involving random error followed by rounding, and we have seen that in this case, depending on parameters, the nonrandom distortion can be benign or problematic.

A lesson from our study is that there are really two candidates for the term "volatility", namely $\langle X, X\rangle_T$, and $\langle f(X), f(X)\rangle_T$, and that in some cases these can diverge substantially. To further investigate what quantity one wishes to estimate, there is need for more research into the use of realized volatility estimates in such applications as portfolio management, options trading and forecasting.

ACKNOWLEDGEMENTS

REFERENCES

Y. Aït-Sahalia, et al. (2005). 'How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise'. *Review of Financial Studies* **18**:351–416.

T. G. Andersen, et al. (2000). 'Great realizations'. *Risk* **13**:105–108.

T. G. Andersen, et al. (2001). 'The Distribution of Exchange Rate Realized Volatility'. *Journal of the American Statistical Association* **96**:42–55.

F. M. Bandi & J. R. Russell (2003). 'Microstructure Noise, Realized Volatility and Optimal Sampling'. Tech. rep., University of Chicago Graduate School of Business.

O. E. Barndorff-Nielsen & N. Shephard (2002). 'Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models'. *Journal of the Royal Statistical Society, B* **64**:253–280.

M. Chernov & E. Ghysels (2000). 'A Study Towards a Unified Approach to the Joint Estimation of Objective and Risk Neutral Measures for the Purpose of Options Valuation'. *Journal of Financial Economics* **57**:407–458.

M. M. Dacorogna, et al. (2001). *An Introduction to High-Frequency Finance.* Academic Press, San Diego.

S. Delattre & J. Jacod (1997). 'A Central Limit Theorem for Normalized Functions of the Increments of a Diffusion Process, in the Presence of Round-Off Errors'. *Bernoulli* **3**:1–28.

J. D. Farmer, et al. (2005). 'A Random Order Placement Model of Price Formation in the Continuous Double Auction'. In L. Blume & S. Durlauf (eds.), *The Economy as an Evolving Complex System, III*, pp. 133–173. Oxford University Press, Oxford, England.

A. R. Gallant, et al. (1999). 'Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance'. *The Review of Economics and Statistics* **81**:617–631.

A. Gloter (2000). *Estimation des Paramètres d'une Diffusion Cachée.* Ph.D. thesis, Université de Marne-la-Vallée.

P. Hall & C. C. Heyde (1980). *Martingale Limit Theory and Its Application.* Academic Press, Boston.

P. R. Hansen & A. Lunde (2006). 'Realized Variance and Market Microstructure Noise'. *Forthcoming in the Journal of Business and Economic Statistics* .

J. Hasbrouck (1996). 'Modeling Market Microstructure Time Series'. In C. R. Rao & G. S. Maddala (eds.), *Handbook of Statistics, volume 14*, pp. 647–692. North-Holland, Amsterdam.

J. Hull & A. White (1987). 'The Pricing of Options on Assets with Stochastic Volatilities'. *Journal of Finance* **42**:281–300.

J. Jacod (1996). 'La Variation Quadratique du Brownien en Présence d'Erreurs d'Arrondi'. *Astérisque* **236**:155–162.

J. Jacod & P. Protter (1998). 'Asymptotic Error Distributions for the Euler Method for Stochastic Differential Equations'. *Annals of Probability* **26**:267–307.

J. Jacod & A. N. Shiryaev (2003). *Limit Theorems for Stochastic Processes.* Springer-Verlag, New York, second edn.

J. Kolassa & P. McCullagh (1990). 'Edgeworth series for lattice distributions'. *Annals of Statistics* **18**:981–985.

J. Large (2005). 'Estimating Quadratic Variation When Quoted Prices Change by a Constant Increment'. Tech. rep., Nuffield College Economics Group Working Paper.

P. A. Mykland & L. Zhang (2006). 'ANOVA for Diffusions'. *Annals of Statistics, forthcoming* **34**:# 4.

P. Protter (2004). *Stochastic Integration and Differential Equations: A New Approach.* Springer-Verlag, New York, second edn.

D. Revuz & M. Yor (1999). *Continuous Martingales and Brownian Motion.* Springer-Verlag, Berlin, Germany, third edn.

Y. Zeng (2003). 'A partially-observed model fo0r micro-movement of asset process with Bayes estimation via filtering for a class of micro-movement models of asset price'. *Mathematical Finance* **13**:411–444.

L. Zhang (2006). 'Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach'. *Bernoulli, forthcoming* –:–.

L. Zhang, et al. (2005). 'A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data'. *Journal of the American Statistical Association* **472**:1394–1411.

B. Zhou (1996). 'High-Frequency Data and Volatility in Foreign-Exchange Rates'. *Journal of Business & Economic Statistics* **14**:45–52.