Available online at www.sciencedirect.com

SciVerse ScienceDirect

Stochastic Processes and their Applications 123 (2013) 2728–2751

stochastic processes and their applications

www.elsevier.com/locate/spa

# Measuring the relevance of the microstructure noise in financial data

Cecilia Mancini

*Dipartimento di Scienze per l'Economia e l'Impresa, Università di Firenze, Italy*

Available online 16 April 2013

**Abstract**

We show that the Truncated Realized Variance (TRV) of a SemiMartingale (SM) converges to zero when observations are contaminated by noise. Under the additive i.i.d. noise assumption, a central limit theorem is also proved. In consequence it is possible to construct a feasible test allowing us to measure, for a given path of a given data generating process at a given observation frequency, the relevance of the noise in the data when we want to estimate the *efficient* process integrated variance $IV$. We thus can optimally select the observation frequency at which we can "safely" use TRV. The performance of our test is verified on simulated data. We are especially interested in the application of the test to financial data, and a comparison conducted with Bandi and Russel (2008) and Ait-Sahalia, Mykland and Zhang (2005) mean square error criteria shows that, in order to estimate IV, in many cases we can rely on TRV for lower observation frequencies than previously indicated when using Realized Variance (RV). The advantages of our method are at least two: on the one hand the underlying model for the efficient data generating process is less restrictive in that jumps are allowed (in the form of an Itô SM). On the other hand our criterion is pathwise, rather than based on an average estimation error, allowing for a more precise estimation of IV because the choice of the optimal frequency is based on the observed path. Further analysis on both simulated and empirical financial data is conducted in Lorenzini (2012) [15] and is also still in progress.



*Keywords:* Semimartingales with jumps; Noisy data; Integrated variance; Threshold estimation; Test to select optimal sampling frequency; Financial asset prices

*E-mail address:* cecilia.mancini@unifi.it.

0304-4149/$ - see front matter 
http://dx.doi.org/10.1016/j.spa.2013.04.003

## 1. Introduction

We can observe noisy data $Y_{t_i} = X_{t_i} + \varepsilon_{t_i}$ of an *efficient* data generating process (DGP) $X$, which is assumed to be an Itô semimartingale (SM) with continuous martingale part $\int \sigma_s dW_s$. If we want to estimate the *integrated variance* $IV \doteq \int_0^T \sigma_s^2 ds$ of $X$, we have to decide whether to use an estimator explicitly accounting for the contribution of the noise process $\varepsilon$ (e.g. by pre-averaging the observations as in [19]) or to directly apply an estimator which is consistent in the absence of noise. This depends on whether the noise is relevant or not in our data, which is determined by the magnitude of $\mathrm{Var}(\varepsilon_{t_i})$ but also by the frequency at which we pick the observations. We are especially interested in the application of our results to financial data, where $X_{t_i}$ represents the logarithm of the *efficient* price of an asset at time $t_i$ and $\varepsilon_{t_i}$, $i = 1..n$ are called *microstructure noises*. Given a time series generated by a Brownian semimartingale (BSM), i.e. a SM without jumps, it is well known that the realized variance ($RV_h$) converges to $IV$, as the observation frequency $h$ tends to 0. If the BSM observations are noisy, we can look at the signature plot (SP) of the realized variance as a function of $h$ to decide whether at a predetermined frequency the noise contamination is relevant or not [8]: when the noise is judged to be negligible, we rely on $RV_h$ as a measure of $IV$. However the observation step $\hat{h}$ visually selected by means of the SP is not necessarily such that $RV_{\hat{h}}$ delivers a reliable estimate of $IV$, given that $RV_{\hat{h}}$ cannot disentangle the estimation error due to the choice of a too large $h$ from the error induced by the presence of the noise. Moreover, in the presence of jumps in the DGP, $RV_h$ undergoes a further source of estimation bias of $IV$, represented by the sum of the squared jumps. Another important criterion used to establish the limit frequency at which the noise can be neglected is theoretically minimizing the conditional mean square estimation error $RV_h - IV$, as described in [4,3,22]. However also in this case $X$ is assumed to have continuous paths. Further, the selected $h$ is optimal on average, along many paths of the price process, while it is possible that the optimal step for a given day is different from the step which is optimal in another day. This makes it useful to have a further tool allowing to establish, for a fixed path of a fixed asset and a given frequency, whether the noise is contaminating the asset returns in a non-negligible way or not. We are thus going to propose a test and to check its reliability on simulated data. The application to empirical financial data has been done in [15] and is also still in progress. Questions that we judge to be interesting are (1) checking whether, as stated by some authors (as in [20]), the mid-quotes are less affected by noise than the transaction prices and at which extent; (2) for a given high frequency, checking how much, when pre-averaging the data, the (normalized) pre-averaged time series has been decontaminated by the noise; (3) for an observation frequency at which the noise is judged to be negligible by our test, comparing the performances of TRV and pre-averaged TRV.

The paper is organized as follows: Section 2 draws the framework we are considering, Section 3 contains the main results allowing to construct out test, Section 4 illustrates how the test is constructed and how it works. In Section 5 we implement the test on simulated data in order to check whether its responses are reliable, meaning that when the test judges the noise to be relevant then the estimation error $\hat{IV}_h - IV$ is high while it is low otherwise. Appendix contains the proofs of all the results stated in this paper.

## 2. Model setup

For a fixed $T \in \mathbb{R}$ let us consider the filtered probability space $\mathcal{S}^0 = (\Omega^0, \mathcal{F}^0, \mathcal{F}^0_{t\in[0,T]}, P^0)$ generated by a Brownian motion $W$ and a Poisson random measure $\mu$ (possibly allowing for