Journal of Econometrics 160 (2011) 190–203

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Edgeworth expansions for realized volatility and related estimators[☆]

Lan Zhang[a], Per A. Mykland[b], Yacine Aït-Sahalia[c,*]

[a] University of Illinois at Chicago, United States
[b] The University of Chicago, United States
[c] Princeton University and NBER, United States

ARTICLE INFO

Article history:
Available online 6 March 2010

JEL classification:
C13
C14
C15
C22

Keywords:
Bias correction
Edgeworth expansion
Market microstructure
Martingale
Realized volatility
Two scales realized volatility

ABSTRACT

This paper shows that the asymptotic normal approximation is often insufficiently accurate for volatility estimators based on high frequency data. To remedy this, we derive Edgeworth expansions for such estimators. The expansions are developed in the framework of small-noise asymptotics. The results have application to Cornish–Fisher inversion and help setting intervals more accurately than those relying on normal distribution.

1. Introduction

Volatility estimation from high frequency data has received substantial attention in the recent literature.[1] A phenomenon which has been gradually recognized, however, is that the standard estimator, realized volatility or realized variance (RV, hereafter), can be unreliable if the microstructure noise in the data is not explicitly taken into account. Market microstructure effects are surprisingly prevalent in high frequency financial data. As the sampling frequency increases, the noise becomes progressively more dominant, and in the limit swamps the signal. Empirically, sampling a typical stock price every few seconds can lead to volatility estimates that deviate from the true volatility by a factor of two or more. As a result, the usual prescription in the literature is to sample sparsely, with the recommendations ranging from five to thirty minutes, even if the data are available at much higher frequencies.

More recently various RV-type estimators have been proposed to take into account the market microstructure impact. For example, in the parametric setting, Aït-Sahalia et al. (2005) proposed likelihood corrections for volatility estimation; in the nonparametric context, Zhang et al. (2005) proposed five different RV-like estimation strategies, culminating with a consistent estimator based on combining two time scales, which we called TSRV (two scales realized volatility).[2]

One thing in common among various RV-type estimators is that the limit theory predicts that the estimation errors of these estimators should be asymptotically mixed normal. Without noise, the asymptotic normality of RV estimation errors dates back to at least Jacod (1994) and Jacod and Protter (1998). When microstructure noise is present, the asymptotic normality of the standard RV estimator (as well as that of the subsequent refinements that are

☆ Financial support from the NSF under grants SBR-0350772 (Aït-Sahalia) and DMS-0204639, DMS 06-04758, and SES 06-31605 (Mykland and Zhang) is gratefully acknowledged. The authors would like to thank the editors and referees for helpful comments and suggestions.

* Corresponding address: Princeton University and NBER, Department of Economics, Princeton, NJ 08544-1021, United States. Tel.: +1 609 258 4015.

E-mail address: yacine@princeton.edu (Y. Aït-Sahalia).

1 See, e.g., Dacorogna et al. (2001), Andersen et al. (2001b), Zhang (2001), Barndorff-Nielsen and Shephard (2002), Meddahi (2002) and Mykland and Zhang (2006).

2 A natural generalization of TSRV, based on multiple time scales, can improve the estimator's efficiency (Zhang, 2006). Also, since the development of the two scales estimators, two other classes of estimators have been developed for this problem: realized kernels (Barndorff-Nielsen et al., 2008, 2011), and pre-averaging (Podolskij and Vetter, 2009; Jacod et al., 2009). Other strategies include Zhou (1996, 1998), Hansen and Lunde (2006), and Bandi and Russell (2008). Studying the Edgeworth expansions of these statistics is beyond the scope of this paper, instead we focus on the statistics introduced by Zhang et al. (2005).

0304-4076/$ – see front matter
doi:10.1016/j.jeconom.2010.03.030

robust to the presence of microstructure noise, such as TSRV) was established in Zhang et al. (2005).

However, simulation studies do not agree well with what the asymptotic theory predicts. As we shall see in Section 5, the error distributions of various RV-type estimators (including those that account for microstructure noise) can be far from normal, even for fairly large sample sizes. In particular, they are skewed and heavy-tailed. In the case of basic RV, such non-normality appears to first have been documented in simulation experiments by Barndorff-Nielsen and Shephard (2005).[3]

We argue that the lack of normality can be caused by the coexistence of a small effective sample size and small noise. As a first-order remedy, we derive Edgeworth expansions for the RV-type estimators when the observations of the price process are noisy. What makes the situation unusual is that the errors (noises) ϵ are very small, and if they are taken to be of order $O_p(1)$, their impact on the Edgeworth expansion may be exaggerated. Consequently, the coefficients in the expansion may not accurately reflect which terms are important. To deal with this, we develop expansions under the hypothesis that the size of $|\epsilon|$ goes to zero, as stated precisely at the beginning of Section 4. We will document that this approach predicts the small sample behavior of the estimators better than the approach where $|\epsilon|$ is of fixed size. In this sense, we are dealing with an unusual type of Edgeworth expansion.

One can argue that it is counterfactual to let the size of ϵ go to zero as the number of observations go to infinity. We should emphasize that we do not literally mean that the noise goes down the more observations one gets. The form of asymptotics is merely a device to generate appropriately accurate distributions. Another problem where this type of device is used is for ARMA processes with nearly unit root (see, e.g. Chan and Wei, 1987), or the local-to-unity paradigm. In our setting, the assumption that the size of ϵ goes down has produced useful results in Sections 2 and 3 of Zhang et al. (2005). For the problem discussed there, shrinking ϵ is the only known way of discussing bias-variance trade-off rigorously in the presence of a leverage effect. Note that a similar use of triangular array asymptotics has been used by Delattre and Jacod (1997) in the context of rounding, and by Gloter and Jacod (2001) in the context of additive error. Another interpretation is that of small-sigma asymptotics, cf. the discussion in Section 4.1 below.

It is worth mentioning that jumps are not the likely causes leading to the non-normality in RV's error distributions in Section 5, as we model both the underlying returns and the volatility as continuous processes. Also, it is important to note that our analysis focuses on normalized RV-type estimators, rather than studentized RV which has more immediate implementation in practice. In other words, our Edgeworth expansion has the limitation of conditioning on volatility processes, while hopefully it sheds some light on how an Edgeworth correction can be done for RV-type estimators while allowing for the presence of microstructure noise. For an Edgeworth expansion applicable to the studentized (basic) RV estimator when there is no noise, one can consult Gonçalves and Meddahi (2009). Their expansion is used for assessing the accuracy of the bootstrap in comparison to the first order asymptotic approach. See also Gonçalves and Meddahi (2008). Edgeworth expansions for realized volatility are also developed by Lieberman and Phillips (2006) for inference on long memory parameters.

With the help of Cornish–Fisher expansions, our Edgeworth expansions can be used for the purpose of setting intervals that are more accurate than the ones based on the normal distribution. Since our expansions hold in a triangular array setting, they can also be used to analyze the behavior of bootstrapping distributions. A nice side result in our development, which may be of use in other contexts, shows how to calculate the third and fourth cumulants of integrals of Gaussian processes with respect to Brownian motion. This can be found in Proposition 4.

The paper is organized as follows. In Section 2, we briefly recall the estimators under consideration. Section 3 gives their first order asymptotic properties, and reports initial simulation results which show that the normal asymptotic distribution can be unsatisfactory. So, in Section 4, we develop Edgeworth expansions. In Section 5, we examine the behavior of our small-sample Edgeworth corrections in simulations. Section 6 concludes. Proofs are in the Appendix.

2. Data structure and estimators

Let $\{Y_{t_i}\}$, $0 = t_0 \le t_1 \le \cdots t_n = T$, be the observed (log) price of a security at time $t_i \in [0, T]$. The basic modelling assumption we make is that these observed prices can be decomposed into an underlying (log) price process X (the signal) and a noise term ϵ, which captures a variety of phenomena collectively known as market microstructure noise. That is, at each observation time t_i, we have

$$Y_{t_i} = X_{t_i} + \epsilon_{t_i}. \tag{2.1}$$

Let the signal (latent) process X follow an Itô process

$$\mathrm{d}X_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}B_t, \tag{2.2}$$

where B_t is a standard Brownian motion. We assume that, μ_t, the drift coefficient, and σ_t^2, the instantaneous variance of the returns process X_t, will be (continuous) stochastic processes. We do not, in general, assume that the volatility process, when stochastic, is orthogonal to the Brownian motion driving the price process.[4] However, we will make this assumption in Section 4.3.

Let the noise ϵ_{t_i} in (2.1) satisfy the following assumption,

$$\epsilon_{t_i} \text{ i.i.d. with } E(\epsilon_{t_i}) = 0, \quad \text{and}$$
$$\mathrm{Var}(\epsilon_{t_i}) = E\epsilon^2. \text{ Also } \epsilon \perp\!\!\!\perp X \text{ process}, \tag{2.3}$$

where $\perp\!\!\!\perp$ denotes independence between two random quantities. Note that our interest in the noise is only at the observation times t_i's, so, model (2.1) does not require that ϵ_t exists for every t. We are interested in estimating

$$\langle X, X\rangle_T = \int_0^T \sigma_t^2 \mathrm{d}t, \tag{2.4}$$

the integrated volatility or quadratic variation of the true price process X, assuming model (2.1), and assuming that Y_{t_i}'s can be observed at high frequency. In particular, we focus on estimators that are nonparametric in nature, and as we will see, are extensions of RV.

Following Zhang et al. (2005), we consider five RV-type estimators. Ranked from the statistically least desirable to the most desirable, we start with (1) the "all" estimator $[Y, Y]^{(\mathrm{all})}$, where RV is based on the entire sample and consecutive returns are used; (2) the sparse estimator $[Y, Y]^{(\mathrm{sparse})}$, where the RV is based on a sparsely sampled returns series. Its sampling frequency is often arbitrary or selected in an ad hoc fashion; (3) the optimal, sparse estimator $[Y, Y]^{(\mathrm{sparse,opt})}$, which is similar to $[Y, Y]^{(\mathrm{sparse})}$ except that the sampling frequency is pre-determined to be optimal in the

[3] We emphasize that the phenomenon we describe is the distribution of the estimation error of volatility measures. This is different from the well known empirical work demonstrating the non-normality of the *unconditional distribution* of RV estimators (see for example Zumbach et al., 1999; Andersen et al., 2001a,b), where the dominant effect is the behavior of the true volatility itself.

[4] See the theorems in Zhang et al. (2005) for the precise assumptions.