

# Inference about Realized Volatility using Infill Subsampling\*

Ilze Kalnina<sup>†</sup> and Oliver Linton<sup>‡</sup>

The London School of Economics

August 31, 2007

## Abstract

We investigate the use of subsampling for conducting inference about the quadratic variation of a discretely observed diffusion process under an infill asymptotic scheme. We show that the usual subsampling method of Politis and Romano (1994) is inconsistent when applied to our inference question. Recently, a type of subsampling has been used to do an additive bias correction to obtain a consistent estimator of the quadratic variation of a diffusion process subject to measurement error, Zhang, Mykland, and Ait-Sahalia (2005). This subsampling scheme is also inconsistent when applied to the inference question above. This is due to a high correlation between estimators on different subsamples. We discuss an alternative approach that does not have this correlation problem; however, it has a vanishing bias only under smoothness assumptions on the volatility path. Finally, we propose a subsampling scheme that delivers consistent inference without any smoothness assumptions on the volatility path. This is a general method and can be potentially applied to conduct inference for quadratic variation in the presence of jumps and/or microstructure noise by subsampling appropriate consistent estimators.

KEYWORDS: Realised Volatility, Semimartingale, Subsampling, Infill Asymptotic Scheme

JEL CLASSIFICATION: C12

---

\*We would like to thank Peter Bühlmann and Nour Meddahi for helpful comments, as well as seminar participants at Oberwolfach (March 19 – 23, 2007). This research was supported by the ESRC and the Leverhulme foundation.

<sup>†</sup>Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.  
E-mail address: [i.kalnina@lse.ac.uk](mailto:i.kalnina@lse.ac.uk)

<sup>‡</sup>Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.  
E-mail address: [o.linton@lse.ac.uk](mailto:o.linton@lse.ac.uk)

# 1 Introduction

Recently, estimation of integrated volatility of the price process has been a fruitful area of research, Barndorff-Nielsen and Shephard (2002). Applications include evaluation of variance forecasting models, portfolio management, and asset pricing. Different estimators of integrated volatility have been proposed depending on the statistical model used. The richer is the model for the price process (e.g., with jumps and/or market microstructure noise added to the diffusion process), the more complicated estimators have been used. Also, if not the estimator itself, its asymptotic (conditional) variance and hence inference changes depending on the assumptions on the price process. For example, Aït-Sahalia, Mykland, and Zhang (2006) and Kalnina and Linton (2006) consider price being a diffusion process contaminated by autocorrelated (former) or endogenous (latter) measurement error. The result is more complicated expressions for the variance of the estimator. These variances cannot be estimated using the methods that work with i.i.d. errors that are independent of the latent price measurement error (or any other method known to us).

The aim of this paper is to investigate the potential of subsampling as a robust method for inference in the pure infill asymptotic framework, which could work in the presence of different assumptions on the observed price process. We choose the relatively simple setup where the underlying price process follows a diffusion, as there are no subsampling results for inference in this framework. Without jumps or measurement error, a consistent estimator for quadratic variation of the price process is the widely used realised volatility. The question we want to answer is, can subsampling be used to conduct inference on realised volatility?

The subsampling method of Politis and Romano (1994) has been shown to be useful in many situations as a way of conducting inference under weak assumptions and without utilizing knowledge of limiting distributions. The basic intuition it builds on goes as follows. Imagine the standard setting of discrete time with long-span (also called increasing domain) asymptotics. Take some general estimator  $\hat{\theta}_n$  (think of i.i.d.  $X_i$ 's,  $\theta = E(X)$ ,  $\hat{\theta}_n = \frac{1}{n} \sum X_i$ ), with

$$\tau_n(\hat{\theta}_n - \theta) \implies N(0, V)$$

as  $n \rightarrow \infty$ , where  $\implies$  denotes convergence in distribution. In this paper we will be interested only in the standard case when  $\tau_n = \sqrt{n}$ . Now, construct subsamples of  $m = m(n)$  consecutive observations, starting at different values (whether they are overlapping or not is irrelevant here), where  $m = m(n) \rightarrow \infty$  as  $n \rightarrow \infty$  but  $m/n \rightarrow 0$ . Then, the asymptotic distribution of  $\tau_m(\hat{\theta}_{n,m,j} - \theta)$

is the same, i.e.,

$$\tau_m \left( \widehat{\theta}_{n,m,j} - \theta \right) \implies N(0, V), \quad j = 1, \dots, K \quad (1)$$

for each subsample. Hence, we can estimate  $V$  by the sample variance of  $\tau_m \widehat{\theta}_{n,m,j}$  (with centering around  $\widehat{\theta}_n$ , the proxy for the true value  $\theta$ ). This yields

$$\widehat{V} = m \times \frac{1}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \widehat{\theta}_n \right)^2, \quad (2)$$

and we have

$$\widehat{V} \xrightarrow{p} V,$$

where  $\xrightarrow{p}$  denotes convergence in probability.

This is our starting point. We show that in an infill sampling scheme a direct application of this method does not achieve the required consistency. The intuition behind this failure is straightforward. We do not have a stationary underlying process (in particular, we have a heteroscedasticity that does not vanish in the limit). Therefore, realised volatility on these subsamples ( $\widehat{\theta}_{n,m,j}$ ) does not converge to quadratic variation over the full time interval ( $\theta$ ). That is, (1) does not hold.

Recently, the word subsampling has been used in connection with the estimation of quadratic variation of a latent price process subject to market microstructure noise, see Zhang, Mykland, and Aït-Sahalia (2005) and Barndorff-Nielsen and Shephard (2007). The subsampling scheme in this setting is slightly different from the usual one and is perhaps better called ‘Infill Price’ subsampling. Zhang, Mykland, and Aït-Sahalia (2005) use this ‘Infill Price’ subsampling to define a bias correction method that achieves consistent estimation. For these type of subsamples, equation (1) holds, so we explore the possibility of their use for inference. We show that ‘Infill Price’ subsampling does not deliver consistent inference for the realised volatility, due to high correlation between subsamples. We recover consistency by modifying the ‘Infill Price’ subsampling in ‘Infill Returns’ subsampling. However, this method requires restrictive assumptions on the volatility path to achieve consistency.

Finally, we propose ‘Subset Centered Infill’ subsampling, which does achieve the required consistency without any smoothness assumptions on the volatility path. It builds on the basic principles of Politis and Romano (1994), but uses a different centering for each subsample.

The remainder of the paper is organized as follows. Section 2 outlines the model and some basic facts about realized volatility. We describe the usual subsampling method and show its inconsistency in section 3. Section 4 introduces ‘Infill Price’ subsampling and shows it is also inconsistent, albeit

asymptotically unbiased. Section 5 introduces ‘Infill Returns’ subsampling and shows its consistency, under some smoothness assumption on the volatility path. Section 6 introduces the ‘Subset Centered Infill’ Subsampling and shows its consistency. Section 7 compares different methods for inference in a Monte Carlo simulation study. Section 8 concludes. We follow the notation of Politis, Romano, and Wolf (1999) for easier comparison with the subsampling literature.

## 2 The Model and Quantities of Interest

Suppose that we have a scalar Brownian diffusion process

$$dX_t = \mu_t dt + \sigma_t dW_t, \quad (3)$$

where  $W_t$  is standard Brownian motion, while the stochastic process  $\mu_t$  is locally bounded and  $\sigma_t$  is càdlàg. Suppose that we observe  $X$  at times  $t_0, \dots, t_n$  on the interval  $[0, T]$ , where  $T$  is fixed, so that our asymptotics are infill (as  $n \rightarrow \infty$ ). Without loss of generality we take the observations equally spaced and  $T = 1$ .

We are primarily interested in the quadratic variation of the process over the observation interval

$$QV_X = \int_0^1 \sigma_s^2 ds,$$

which is a random variable depending on the realization of the volatility path  $\{\sigma_t, t \in [0, 1]\}$ . The usual estimator of  $QV_X$  is the realized volatility

$$RV_n = \sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})^2.$$

This satisfies

$$\begin{aligned} \sqrt{n}(RV_n - QV_X) &\Longrightarrow MN(0, V) \\ V &= 2IQ = 2 \int_0^T \sigma_s^4 ds, \end{aligned} \quad (4)$$

where  $MN(0, V)$  denotes a mixed normal distribution with random conditional variance  $V$  independent of the underlying normal distribution, that is, the limiting pdf is of the form  $f(x) = \int \phi_{0,v}(x) f_V(v) dv$ , where  $f_V$  denotes the pdf of  $V$  and  $\phi_{0,v}(x) = \exp(-x^2/2v^2)/\sqrt{2\pi v}$ . The convergence (4) follows from Barndorff-Nielsen and Shephard (2002), and is stable in law, Aldous and Eagleson

(1978), meaning that the convergence holds jointly with the random variable  $V$ . Barndorff-Nielsen and Shephard (2002) also show the "feasible" CLT

$$\tilde{V}^{-1/2}\sqrt{n}(RV_n - QV_X) \implies N(0, 1), \quad (5)$$

where  $\tilde{V} = 2IQ_n$  is a consistent estimator of  $V = 2IQ$  in the sense that  $\tilde{V}/V \xrightarrow{p} 1$ . Here,  $IQ_n$  is the realized quarticity,

$$IQ_n = \frac{n}{3} \sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})^4,$$

which successfully mimics the structure of  $\int_0^T \sigma_s^4 ds$ . The result (5) allows the construction of consistent confidence intervals for  $QV_X$ . For example, a two-sided level  $\alpha$  interval is given by  $\tilde{\mathcal{C}}_\alpha = RV_n \pm z_{\alpha/2} \tilde{V}^{1/2} / \sqrt{n}$ , where  $z_\alpha$  is the  $\alpha$  quantile from a standard normal distribution, and this has the property that  $\Pr[QV_X \in \tilde{\mathcal{C}}_\alpha] \rightarrow 1 - \alpha$ . They also propose to construct intervals from the limiting distribution of  $\ln RV_n$ , which would thereby respect the positivity of  $QV_X$  by giving an asymmetric interval. Mykland and Zhang (2006,7) have proposed alternative estimators of  $V$  that are more efficient than  $\tilde{V}$  under the sampling scheme (3) and can also be used to construct intervals based on the studentized limit theory.

Recently, Goncalvez and Meddahi (2005) have proposed a bootstrap algorithm. They use the i.i.d. bootstrap applied to returns, that is,  $r_{t_i} = X_{t_i} - X_{t_{i-1}}$  are resampled with replacement. In the no-leverage case, i.e., where the stochastic processes  $\{\sigma_t, t \in [0, 1]\}$  and  $\{W_t, t \in [0, 1]\}$  are mutually independent (and  $\mu_t \equiv 0$ ), stock returns are mutually independent, although heterogeneous, conditional on the volatility path. They have shown that this resampling scheme is consistent in the no leverage case. They also proposed a modification called the wild bootstrap, Horowitz (2001), that at least allows for heterogeneity in returns. This method is not only consistent but even achieves some higher order refinements in the case of no leverage. In the leverage case where the processes  $\{\sigma_t, t \in [0, 1]\}$  and  $\{W_t, t \in [0, 1]\}$  are no longer independent, stock returns are dependent over time both unconditionally and conditionally on the volatility path. In a long span setting, it would generally be fatal to ignore the dependence structure present in the leverage case in this way.<sup>1</sup> Nevertheless, since the distribution theory for integrated volatility is the same under leverage as under no leverage it may be that the i.i.d. bootstrap is consistent for the studentized statistic as

---

<sup>1</sup>As Horowitz (2001) says, "Bootstrap sampling must be carried out in a way that suitably captures the dependence of the data-generation process".

their simulations suggest.<sup>2</sup>

The purpose of this paper is to explore the use of subsampling as a means of conducting inference without utilizing the knowledge of the asymptotic (conditional) variance of the estimator, that is we do not work with a studentized statistic, which requires a consistent estimator of the asymptotic conditional variance. Therefore, it can have the potential to be applied with different underlying assumptions, which lead to different, possibly very complicated, asymptotic (conditional) variances. For example, when there are jumps in (3), the limiting distribution (4) no longer holds, as the random conditional variance changes to (10). Unlike the bootstrap method of Goncalvez and Meddahi (2005), the subsampling method itself does not impose or exploit the conditional independence structure of the no-leverage case, and in other contexts is known to adapt to whatever dependence structure is present since the data is used in blocks. It also requires no random number generator and can have computational advantages. The subsampling method should also be robust to non-equally spaced data and to heavy tails, Politis and Romano (1994).<sup>3</sup>

On the other hand, when all the conditions of Barndorff-Nielsen and Shephard (2002) hold, their asymptotic studentization or the Goncalvez and Meddahi (2005) bootstrap method are likely to perform much better, which is not surprising since they make explicit use of that structure, see Horowitz (2001).

We maintain the assumption of no leverage, i.e., we assume independence of the stochastic processes  $\{\mu_t, \sigma_t, t \in [0, 1]\}$  and  $\{W_t, t \in [0, 1]\}$ . For the "negative results", Proposition 1 and 2, this is of course no loss of generality. The implication of this assumption is that one can treat  $V$  as fixed; we therefore just consider estimation of  $V$ . We first show that several standard approaches to subsampling do not work under the infill asymptotics. We then establish that various modifications do provide consistent inference. Although we currently use the no leverage assumption for our proofs, our Monte Carlo simulations indicate that our proposed subsampling method should also work in the absence of this restriction. In the sequel, we work with the conditional distribution given  $\{\mu_t, \sigma_t, t \in [0, 1]\}$ . Also, we denote  $QV_X$  by  $\theta$ .

---

<sup>2</sup>This result is not unlike what happens in kernel density estimation for stationary weakly dependent processes, Horowitz (2001). The limiting distribution of the kernel estimator is as if the data were i.i.d. from a population with the same marginal distribution. Also, it is known that the i.i.d. bootstrap will also provide consistent inference in this case.

<sup>3</sup>This latter issue might be relevant in the context of Fractional Brownian Motion and related heavy tailed processes, Samorodnitsky and Taqqu (1994).

### 3 Regular subsampling

We start our investigation of subsampling for the infill asymptotic framework with the well established method of Politis and Romano (1994), which is developed for the long span asymptotic framework.

In Politis, Romano, and Wolf (1999) there is some discussion about subsampling for continuous parameter random fields, which includes continuous time processes as a special case. Unfortunately, the discussion is suited to the continuous observation and increasing domain case where  $T \rightarrow \infty$ . Also, they assume strong stationarity, which we do not have.<sup>4</sup> They use the following notation, which we have specialized for the scalar parameter case.<sup>5</sup> Let

$$E_{j,m} = \{t_i : j - 1 < i \leq j - 1 + m\}$$

$$Y_j = \{X(t) : t \in E_{j,m}\}$$

$$\hat{\theta}_{n,m,j} = \hat{\theta}(Y_j),$$

where  $\hat{\theta}(Y_j)$  is the estimator computed with the data  $Y_j$ , in particular

$$\hat{\theta}(Y_j) = \sum_{t_i \in E_{j,m}} (X_{t_i} - X_{t_{i-1}})^2.$$

We have  $0 < j \leq K$ , where  $K$  is the number of subsamples,  $K = n - m + 1$ . See Figure 1 in the appendix for a graphical illustration of the subsamples. In our case where  $X$  is only observed at a discrete set of points the data  $Y_j$  consists of a finite set of points of lower cardinality than  $n$ . The estimator  $\hat{\theta}_n = \hat{\theta}(Y) = RV_n$ , when  $Y$  consists of all data.

Assumption 5.3.1 of Politis, Romano and Wolf (1999) is satisfied, i.e., the sampling distribution of  $\tau_n(\hat{\theta}_n - \theta)$  converges weakly. Therefore, in the setting of continuous observations and long-span asymptotics, the intuition laid out in the introduction applies and  $V$  should be approximated by (2), i.e.,

$$\hat{V}_{regular} = m \times \frac{1}{K} \sum_{j=1}^K \left( \hat{\theta}_{n,m,j} - \hat{\theta}_n \right)^2.$$

---

<sup>4</sup>Subsampling results have also been developed for heteroscedasticity that vanishes in the limit, i.e., is close to homoscedasticity. This is an unrealistic setting in infill asymptotics.

<sup>5</sup>We will present the version of regular subsampling that uses overlapping subsamples. All the results hold also with non-overlapping subsamples. However, for the cases discussed in Politis, Romano, and Wolf (1999), it achieves higher efficiency.

However, in our setting, it is easy to see that  $\widehat{V}$  is not a consistent estimator for  $V$ . This is due to  $\widehat{\theta}_{n,m,j}$  only estimating a part of  $\widehat{\theta}_n$  and the asymptotic distribution of  $\tau_m(\widehat{\theta}_{n,m,j} - \theta)$  diverging to infinity instead of being the same as that of  $\tau_n(\widehat{\theta}_n - \theta)$ . As a result, we have

PROPOSITION 1. *Let  $m \rightarrow \infty$  and  $m/n \rightarrow 0$  as  $n \rightarrow \infty$ . We have*

$$\mathbb{E} \left( \widehat{V}_{regular} \right) = m\theta^2 + o(m).$$

## 4 Infill price subsampling

In this and the next section we consider two subsampling schemes that are motivated by the literature on estimating QV in the presence of measurement error. Both recover the principle that  $\tau_m(\widehat{\theta}_{n,m,j} - \theta)$  has the same asymptotic distribution as  $\tau_n(\widehat{\theta}_n - \theta)$ .

First, we explore a subsampling scheme that exactly mirrors the construction of subsamples in the estimator of  $QV_X$  of Zhang, Mykland, and Ait-Sahalia (2005). Let the subsamples consist of data that are  $K$  observations apart. In particular, write  $K \times (m + 1) = n$  and construct the  $j^{th}$  subsample  $Y_j$  and the estimator of  $\theta$  on this subsample as follows (see Figure 3 in the appendix for a graphical illustration),

$$E_{j,m} = \{t_s : s = j + i(K + 1), i = 1, \dots, m\},$$

$$Y_j = \{X(t) : t \in E_{j,m}\},$$

$$\widehat{\theta}_{n,m,j} = \widehat{\theta}(Y_j).$$

In each subsample we have  $m + 1$  (log-)prices and hence  $m$  returns. For  $j = 1, \dots, K$

$$\widehat{\theta}_{n,m,j} = \sum_{t \in E_{j,m}} (X_t - X_{t,-})^2 = \sum_{i=1}^m \left( X_{t_{j+iK}} - X_{t_{j+(i-1)K}} \right)^2,$$

where  $X_{t,-}$  denotes the preceding element to  $X_t$  where  $X_t \in E_{j,m,h}$ .

We can easily see that  $\tau_m(\widehat{\theta}_{n,m,j} - \theta)$  has the same asymptotic distribution as  $\tau_n(\widehat{\theta}_n - \theta)$ . Therefore, we consider the estimator of  $V = 2IQ$  as in (2),

$$\widehat{V}_{\text{Infill price}} = m \times \frac{1}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \widehat{\theta}_n \right)^2.$$

PROPOSITION 2. Let  $m \rightarrow \infty$  and  $m/n \rightarrow 0$  as  $n \rightarrow \infty$ . We have

$$\mathbb{E} \left( \widehat{V}_{\text{Infill price}} \right) \rightarrow \mathbb{E}V, \quad \text{Var} \left( \widehat{V}_{\text{Infill price}} \right) = O(1). \quad (6)$$

This estimator is asymptotically unbiased, but it never converges in probability to the true  $V$ . Why do we have this problem? Note that any two subsamples fully overlap (in terms of time covered by variances involved; see Figure 3), apart from end effects. As a result, the asymptotic correlation between estimators on any two subsamples is one. We have

$$\text{Cov} \left( \widehat{\theta}_{n,m,i}^2, \widehat{\theta}_{n,m,j}^2 \right) \sim \text{Var} \left( \widehat{\theta}_{n,m,i}^2 \right) = O(m^{-2}),$$

and so the failure of consistency follows. We next consider a modification that is motivated by this problem. In the following section we construct subsamples that do not overlap at all while preserving the property that  $\tau_m(\widehat{\theta}_{n,m,j} - \theta)$  has the same asymptotic (conditional) variance as  $\tau_n(\widehat{\theta}_n - \theta)$ .

## 5 Infill Returns Subsampling

To avoid the problem of infill price subsampling, consider subsampling one-period returns  $r_{t_i} = X_{t_i} - X_{t_{i-1}}$  instead of log-prices  $X_t$ . See Figure 4 in the appendix for a graphical illustration.

Let the subsamples consist of returns that are  $K$  observations apart. In particular, write  $K \times (m+1) = n$  and construct the  $j^{\text{th}}$  subsample  $Y_j$  and the estimator of  $\theta$  on this subsample as follows,

$$E_{j,m} = \{t_s : s = j + i(K+1), i = 1, \dots, m\}$$

$$Y_j = \{r(t) : t \in E_{j,m}\}.$$

In each subsample we have  $m$  returns. For  $j = 1, \dots, K$

$$\widehat{\theta}_{n,m,j} = K \sum_{t \in E_{j,m}} r_t^2 = K \sum_{i=1}^m r_{t_{j+(i-1)K}}^2 = K \sum_{i=1}^m \left( X_{t_{j+(i-1)K}} - X_{t_{j+(i-1)K-1}} \right)^2.$$

Define  $\widehat{\theta}_n$  as before and

$$\widehat{V}_{\text{Infill returns}} = m \times \frac{1}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \widehat{\theta}_n \right)^2. \quad (7)$$

As opposed to the previous formulation, we now have  $\text{Cov}(\widehat{\theta}_{n,m,i}, \widehat{\theta}_{n,m,j}) = 0$  for  $i \neq j$ ,<sup>6</sup> which leads us to

PROPOSITION 3. *Let  $m \rightarrow \infty$  and  $m/n \rightarrow 0$  as  $n \rightarrow \infty$ . Assume that volatility paths are a.s. Hölder continuous of order larger than  $1/2$ . We have*

$$\widehat{V}_{\text{Infill returns}} \xrightarrow{p} V. \quad (8)$$

Without imposing some smoothness on the volatility path  $\widehat{V}_{\text{Infill returns}}$  is biased due to the large gaps we have in the subsamples (see Figure 3). For asymptotic unbiasedness of  $\widehat{V}_{\text{Infill returns}}$ , we need the expected value of the estimator on each subsample to converge to  $\theta$  sufficiently fast. If volatility path is càdlàg, we have, conditional on  $\{\sigma_t\}$ ,

$$E\widehat{\theta}_{n,m,j} = K \sum_{i=1}^m \int_{[j+(i-1)K-1]/n}^{[j+(i-1)K]/n} \sigma^2(u) du \longrightarrow \int_0^1 \sigma^2(u) du = \theta$$

due to Riemann integrability of sample paths of  $\{\sigma_t, t \in [0, 1]\}$ . However, what we need for asymptotic unbiasedness of  $\widehat{V}_{\text{Infill returns}}$ , is  $E(\widehat{\theta}_{n,m,j}) - \theta = o(m^{-1/2})$ . This is because of the scaling factor  $m$  in (7), which appears because we are estimating the asymptotic (conditional) variance.

Finally, we consider constructing an estimator of  $V$  that similarly builds on the principle of regular subsampling, but uses a different centering.

## 6 Subset Centered Infill Subsampling

In regular subsampling, the problem was that we were centering our sample variance at "the wrong quantity". In the formula for  $\widehat{V}_{\text{regular}}$ ,

$$\widehat{V}_{\text{regular}} = m \times \frac{1}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \widehat{\theta}_n \right)^2,$$

the quantity  $\widehat{\theta}_n$  plays the role of  $\theta$ , but the problem is that  $\widehat{\theta}_{n,m,j} \not\rightarrow \theta$  and so  $\widehat{V}_{\text{regular}}$  explodes. In the two previous sections we redefined  $\widehat{\theta}_{n,m,j}$  so as to recover the principle  $\widehat{\theta}_{n,m,j} \rightarrow \theta$  and saw this does not work very well. Instead, consider centering estimators at  $\theta_j$  such that  $\widehat{\theta}_{n,m,j} \rightarrow \theta_j$  and then use

---

<sup>6</sup>This covariance is exactly zero if the drift is zero. It is negligible if the drift is non-zero.

the property of integrated variance that it can be added up over subsamples to recover the integrated variance over the full sample,  $\sum_j \theta_j = \theta$ . Now,  $\theta_j$  is not observable and the best proxy for  $\theta_j$  we have is realised volatility over the subsample,  $\hat{\theta}_{n,m,j}$ . Therefore, we have to use something else to play the role of the estimator. So define  $\hat{\theta}_{n,m,J,j}$  to be realised volatility calculated using every  $J^{\text{th}}$  observation of the  $j^{\text{th}}$  subsample of length  $m$  ( $J \ll m$ ), see Figure 2 in the Appendix for a graphical illustration. Since it has a slower rate of convergence than  $\hat{\theta}_{n,m,j}$ , the error of using  $\hat{\theta}_{n,m,j}$  instead of  $\theta$  is negligible. Our estimator of  $V$  becomes

$$\hat{V}_{SC} = \frac{n^2}{mKJ} \sum_{k=1}^K \left( \hat{\theta}_{n,m,J,k} - \hat{\theta}_{n,m,k} \right)^2.$$

Here, the number of subsamples  $K$  depends on how much different subsamples overlap. Increasing the amount of overlap does not change the rate of convergence of  $\hat{V}$ , but it decreases the asymptotic (conditional) variance, so maximum overlap estimator is preferred. See the Monte Carlo simulations (Figure 5e) for an illustration.

PROPOSITION 4. *Let  $m \rightarrow \infty$ ,  $J \rightarrow \infty$ ,  $m/n \rightarrow 0$ , and  $J/m \rightarrow 0$  as  $n \rightarrow \infty$ . We have*

$$\hat{V}_{SC} \xrightarrow{p} V. \tag{9}$$

The estimator is similar in structure to Lahiri, Kaiser, Cressie, and Hsu (1999). They similarly use two grids for subsampling to predict stochastic cumulative distribution function. However, they assume that the underlying process is stationary and their asymptotic framework is mixed infill and increasing domain.

## 7 Numerical Work

In this section we examine the numerical properties of our estimators of  $V$  plus several benchmarks. In our experience the performance of the corresponding confidence intervals here is closely matched by the performance of the estimated variance. In Section 7.1 we simulate the Heston (1993) model with continuous price sample paths. In Section 7.2. we allow price paths to exhibit jumps.

## 7.1 Continuous prince paths

We do a simulation comparison of the above subsampling methods, plus realized quarticity. We simulate the Heston (1993) model:

$$\begin{aligned}dX_t &= (\mu_t - v_t/2) dt + \sigma_t dB_t \\dv_t &= \kappa(\theta - v_t) dt + \gamma v_t^{1/2} dW_t,\end{aligned}$$

where  $v_t = \sigma_t^2$ , and  $B_t, W_t$  are independent standard Brownian motions.

We take parameters from Zhang, Mykland, and Ait-Sahalia (2005):  $\mu = 0.05$ ,  $\kappa = 5$ ,  $\theta = 0.04$ ,  $\gamma = 0.5$ . We set the length of the sample path to 22500, which is a proxy for 23400 that is the number of seconds in a business day. We set the time between observations corresponding to one second when a year is one unit, and the number of replications to be 100,000. For  $\widehat{V}_{\text{regular}}$ ,  $\widehat{V}_{\text{Infill price}}$ ,  $\widehat{V}_{\text{Infill returns}}$ , and  $\widehat{V}_{SC}$ , we use  $m = \sqrt{n}$ . For  $\widehat{V}_{SC}$  we use  $J = 15$ . We hold the volatility sample path constant across simulations for easier comparison. Appendix contains Figure 5 showing the results.

Table 1. Finite sample distributions of different methods

Figure 5a	The volatility sample path
Figure 5b	Kernel density over simulations of $\widehat{V}_{\text{regular}}$
Figure 5c	Kernel density over simulations of $\widehat{V}_{\text{Infill price}}$
Figure 5d	Kernel density over simulations of $\widehat{V}_{\text{Infill returns}}$
Figure 5e	Kernel density over simulations of $\widehat{V}_{SC}$
Figure 5f	Kernel densities over simulations of $\widehat{V}_{\text{Infill price}}$ , $\widehat{V}_{\text{Infill returns}}$ , $\widehat{V}_{SC}$ , and $2IQ_n$

From the simulated volatility sample path (Figure 5a) we can approximate the true  $Var(RV) = V = 2IQ$  and we get  $4.05 \times 10^{-6}$ . Also, this can be approximated by the (scaled) sample variance of RV over simulations. Here is a brief summary comparing the means of feasible methods with infeasible benchmarks:

Table 2. Finite sample means of different methods

Theoretical $\sqrt{2IQ}$	0.00201	infeasible
Finite sample $\sqrt{\text{Var}(RV)/\Delta}$	0.00202	infeasible
Square root of mean over simulations of $\widehat{V}_{\text{regular}}$	0.01720	feasible
Square root of mean over simulations of $\widehat{V}_{\text{Infill price}}$	0.00200	feasible
Square root of mean over simulations of $\widehat{V}_{\text{Infill returns}}$	0.00201	feasible
Square root of mean over simulations of $\widehat{V}_{SC}$ (maximum overlap)	0.00194	feasible
Square root of mean over simulations of $2IQ_n$	0.00201	feasible

We see that  $\widehat{V}_{\text{regular}}$  overestimates the true  $V$  by a large factor, thus supporting the asymptotic result. The other feasible methods work well. We can also see that  $\widehat{V}_{\text{Infill price}}$  is asymptotically unbiased, whereas from Figure 5c in the appendix we see that its distribution over simulations is strongly right-skewed. From Figure 5f we can see that  $2IQ_n$  has the smallest variance, reflecting the fact that it has much faster rate of convergence. Our proposed estimator  $\widehat{V}_{SC}$  has some finite sample negative bias. One could do a finite sample correction and use  $\frac{J}{J-1}\widehat{V}_{SC}$  (recall  $J = 15$ , so this factor is non-negligible), see section A.4.2. for theoretical justification. Adjusted estimator  $\frac{J}{J-1}\widehat{V}_{SC}$  has a smaller negative bias in simulations.

One of the parameters to choose in our proposed estimator  $\widehat{V}_{SC}$  is the amount of overlap between different subsamples, which affects the total number of subsamples  $K$ . We do a simulation exercise to see how exactly it affects the finite sample properties. Figure 5e shows the finite sample distribution for three different scenarios, no overlap, maximum overlap and an intermediate case. We know from theory that the amount of overlap does not affect the convergence rate. The simulations, however, indicate that it decreases the asymptotic conditional variance, i.e.,  $\widehat{V}_{SC}$  with the maximum amount of overlap (with  $K = n - m + 1$ ) gives the lowest variance. This phenomenon is also observed in the long span asymptotic framework.

## 7.2 Jumps

Now we consider the case of possibly discontinuous sample paths. In this case, the realised volatility again converges to the quadratic variation. However, quadratic variation does not coincide with integrated volatility, but instead, it is equal to integrated volatility plus the sum of squared jumps,

$$QV = IV + \sum (X_t - X_{t-})^2,$$

where  $X_{t-}$  is the left limit of  $X_t$ ,

$$X_{t-} = \lim_{s \uparrow t} X_s.$$

We have that  $X_{t-} = X_t$  if there are no jumps at time  $t$ . The asymptotic distribution of  $RV$  is mixed normal with asymptotic (conditional) variance

$$V = 2IQ + 4 \sum (X_t - X_{t-})^2 \sigma_t^2. \quad (10)$$

It is well known that realised quarticity is inconsistent for  $V$ , Barndorff-Nielsen and Shephard (2006). There are two available consistent estimators of  $V$  under the scenario of jumps. One is Aït-Sahalia and Jacod (2005), which uses truncated power variations. The second is that of Veraart (2007) who obtains an estimator for  $V$  as a linear combination of known estimators of  $IQ$  and generalized bipower variation. This estimator entails estimation of  $\sigma_t^2$  for each  $t$  in  $[0,1]$  using a histogram approach. In this section we will consider all estimators of section 7.1 plus Veraart's estimator.

The continuous part of the price paths are simulated as in Heston (1993), where every simulation has its own volatility path. We then add jumps as follows. We consider nine scenarios with none, one, or two jumps per day, with jump times that are uniformly distributed over the day with different jump size distributions (as in, for example, Barndorff-Nielsen and Shephard (2006)). We consider four different sizes of the jumps, governed by a parameter  $p$ . In particular, we draw the size of the jump from a normal p.d.f. with variance  $p$  times the integrated volatility,  $N(0, pIV)$ . This is the setup for jumps considered by Huang and Tauchen (2005) and Veraart (2007).

Table 3. Jump parameterization

scenario	1	2	3	4	5	6	7	8	9
number of jumps per day	0	1				2			
$p$	0	0.1	0.2	0.5	0.7	0.1	0.2	0.5	0.7

We report the results by graphing the kernel densities of the studentised  $RV$  by the infeasible  $QV$  and respective estimator of  $V$ , as follows

$$t = \sqrt{n} \left( \frac{RV - QV}{\sqrt{\widehat{V}}} \right).$$

Ideally, these kernel densities should coincide with the p.d.f of a standard normal distribution, which we superimpose on the estimated density of  $t$ . Figures (6) to (11) contain the results. Nine sub-plots

in each graph represent 9 scenarios as in Table 3. Note that scenario 1 is very similar to the setup of section 7.1, but now we use a different way of representing results (because there is no one true  $V$  anymore, it changes across simulations).

After inspecting Figures (6) to (11) we can make several observations. First, from Figures 6 and 10,  $\widehat{V}_{SC}$  and  $\widehat{V}_{Veraart}$  seem to be good estimators of  $V$  in all scenarios. Most probably, this indicates that  $\widehat{V}^{SC}$  preserves consistency even in this richer context. This seems reasonable, because  $\widehat{\theta}_{n,m,J,k}$  and  $\widehat{\theta}_{n,m,k}$  both converge to the same quantity, i.e., quadratic variation over the (little) time interval they cover.

Second, from Figure 7, we see that  $\widehat{V}_{\text{infill returns}}$  estimator, which was consistent for the variance of  $RV$  in the case of no jumps, is not consistent anymore. The intuition for this failure is straightforward. Although  $\widehat{\theta}_n$  converges to quadratic variation, its subsampled version  $\widehat{\theta}_{n,m,j}$  does not, hence the estimator  $\widehat{V}_{\text{infill returns}}$  overestimates  $V$  by a large factor (hence the peaks at zero of estimated density of  $t_{\text{infill returns}}$ ). This intuition is the same as for the failure of consistency of  $\widehat{V}_{\text{regular}}$  in the case of no jumps (see Section 3).

Third, from Figure 8 we see that the performance of  $\widehat{V}_{\text{infill price}}$  is not worsened by adding jumps. The intuition is that  $\widehat{V}_{\text{infill price}}$  does not suffer from the problem of  $\widehat{V}_{\text{infill returns}}$ . In  $\widehat{V}_{\text{infill price}}$ , both  $\widehat{\theta}_n$  and  $\widehat{\theta}_{n,m,j}$  pick up the same jumps and converge to the same quantity. Therefore, the estimator most probably remains asymptotically unbiased, though retains the problem of high correlation between subsamples and hence is inconsistent. Although a direct comparison between Figures (5c) and (8) should be avoided since the latter does not hold volatility path constant across simulations, we can see that the large mass below the true quantity  $V$  in Figure (5c) is reflected in heavy tails of kernel density in all nine scenarios of (8).

Fourth, from Figure 9, we see that  $2IQ_n$  estimates  $V$  well when there are no jumps, but is clearly not consistent for  $V$  when jumps are present.

## 8 Conclusions and Extensions

In this paper we have investigated the use of subsampling for conducting inference about quadratic variation. We have established two negative results. First, the usual subsampling method as in Politis, Romano, and Wolf (1994) is inconsistent. Second, the subsampling method of Zhang, Mykland, and Aït-Sahalia (2005) is also inconsistent. We have also proposed two alternative subsampling meth-

ods and established their consistency under weak assumptions (given the basic framework we have adopted). The simulation experiments confirm that our methods can be consistent in the presence of leverage and in the presence of jumps.

There is much further work that can be done. Our method can also be applied to other estimators like bipower variation, Barndorff-Nielsen and Shephard (2006) and the leverage estimator of Mykland and Zhang (2007). It is important to generalize the process that  $X$  can follow. For example, one could allow for leverage and jumps. Our simulations show that this can be done to conduct consistent inference for quadratic variation. Inference for integrated volatility would involve subsampling of tripower quarticity or other consistent estimators of integrated volatility. Also, one could allow for measurement error that contaminates  $X$ , in which case one has to consider subsampling  $TSRV$  or realized kernels instead of RV as estimators for  $QV_X$ . Examples of cases where only the asymptotic theory has been developed, but not the tools for inference, include estimation of quadratic variation with autocorrelated measurement error (Aït-Sahalia, Mykland, and Zhang 2006a) or with endogenous measurement error (Kalnina and Linton 2006).

## A Appendix

Recall that we are assuming no leverage, so that we can condition on the volatility path. Also, all arguments are done for a zero drift. Extension to nonzero drift is straightforward (given that we assume a locally bounded drift, and hence bounded drift w.l.o.g. for the purposes of consistency). In the proofs of proposition 3 and 4 we use the following result, which is a modification of Linton (2000, Lemma 1).

LEMMA 0. *Let  $(\theta_n, \sigma)$  be a sequence of random variables with  $\theta_n$  scalar and  $\sigma = \{\sigma_t, t \in [0, 1]\}$ . Suppose that  $E(\theta_n|\sigma) = m_n(\sigma)$  and  $\text{Var}(\theta_n|\sigma) = v_n(\sigma)$  almost surely, where  $m_n(\sigma), v_n(\sigma) \xrightarrow{p} 0$ . Then,  $\theta_n \xrightarrow{p} 0$ .*

### A.1 Proof of Proposition 1

The  $j^{\text{th}}$  subsample contains observations  $\{X_{j-1}, X_j, \dots, X_{m+j-1}\}$ . Then

$$\hat{\theta}(Y_j) = \sum_{i=1}^m (X_{j+i-1} - X_{j+i-2})^2.$$

In this case, the number of subsamples is  $K = n - m + 1$ . Let  $z_{j,i} = \int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u$  and  $y_{j,i} = (\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u)^2 - \sigma_{j,i}^2$ , which are independent (across  $i$  for given  $j$ ) and mean zero random variables conditional on the process  $\sigma_u^2$ , where

$$\sigma_{j,i}^2 = \mathbb{E}\left[\left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u\right)^2\right] = \int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du = O(1/n).$$

In above we use local boundedness of  $\sigma_u^2$  (which follows from the assumption that paths of  $\sigma$  are càdlàg) to conclude orders of magnitude, i.e.,

$$\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du \leq \frac{1}{n} \sup_u \sigma_u^2 = O(1/n).$$

We use this argument of boundedness to conclude stochastic orders of magnitude throughout the appendix.

We have

$$\begin{aligned} \text{var}(z_{j,i}) &= \mathbb{E}\left[\left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u\right)^2\right] = \sigma_{j,i}^2 \\ \text{var}(y_{j,i}) &= \mathbb{E}\left[\left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u\right)^4\right] - \mathbb{E}^2\left[\left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u dW_u\right)^2\right] \\ &= 3\left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du\right)^2 - \left(\int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du\right)^2 \\ &= O(1/n^2). \end{aligned}$$

We show the following lemmas.

LEMMA 1.1. As  $n \rightarrow \infty$

$$\mathbb{E}\left[\sum_{j=1}^K \widehat{\theta}_{n,m,j}\right] = m \int_0^1 \sigma_u^2 du + O\left(\frac{m^2}{n}\right).$$

LEMMA 1.2. As  $n \rightarrow \infty$

$$\sum_{j=1}^K \mathbb{E}\left(\widehat{\theta}_{n,m,j}^2\right) = O\left(\frac{m^2}{n}\right).$$

Then,

$$\begin{aligned}
\widehat{V} &= m \times \frac{1}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \widehat{\theta}_n \right)^2 \\
&= \frac{m}{K} \sum_{j=1}^K \left( \left( \widehat{\theta}_{n,m,j} - \theta \right)^2 + \left( \theta - \widehat{\theta}_n \right)^2 + 2 \left( \widehat{\theta}_{n,m,j} - \theta \right) \left( \theta - \widehat{\theta}_n \right) \right) \\
&= \frac{m}{K} \sum_{j=1}^K \left( \widehat{\theta}_{n,m,j} - \theta \right)^2 + O_p \left( m^{1/2} n^{-1/2} \right) \\
&= \frac{m}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j}^2 + m\theta^2 - 2m\theta \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} + o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\widehat{V}) &= \frac{m}{K} \sum_{j=1}^K \mathbb{E}(\widehat{\theta}_{n,m,j}^2) + m\theta^2 - 2m\theta \frac{1}{K} \sum_{j=1}^K \mathbb{E}(\widehat{\theta}_{n,m,j}) + o(1) \\
&= \frac{m}{K} O\left(\frac{m^2}{n}\right) + m\theta^2 - 2m\theta \frac{1}{K} \left( m \int_0^1 \sigma_u^2 du + O\left(\frac{m^2}{n}\right) \right) + o(1) \\
&= m\theta^2 + o(m) \rightarrow \infty,
\end{aligned}$$

using  $K = n - m + 1$  and  $m^2/n = o(1)$ . ■

PROOF OF LEMMA 1.1. *We have*

$$\begin{aligned}
\mathbb{E} \left[ \sum_{j=1}^K \widehat{\theta}_{n,m,j} \right] &= \sum_{j=1}^K \sum_{i=1}^m \mathbb{E} [z_{j,i}^2] \\
&= \sum_{j=1}^K \sum_{i=1}^m \int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du \\
&= \sum_{j=1}^K \int_{(j-1)/n}^{(m+j-1)/n} \sigma_u^2 du \\
&= m \int_0^1 \sigma_u^2 du + O\left(\frac{m^2}{n}\right).
\end{aligned}$$
■

PROOF OF LEMMA 1.2. *We have*

$$\begin{aligned}
\sum_{j=1}^K \mathbb{E} \left( \widehat{\theta}_{n,m,j}^2 \right) &= \sum_{j=1}^K \mathbb{E} \left[ \left( \sum_{i=1}^m z_{j,i}^2 \right)^2 \right] \\
&= \sum_{j=1}^K \mathbb{E} \left[ \left( \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) \right)^2 \right] \\
&= \sum_{j=1}^K \sum_{i=1}^m \mathbb{E} y_{j,i}^2 + \sum_{j=1}^K \sum_{i=1}^m \sum_{i'=1}^m \sigma_{j,i}^2 \sigma_{j,i'}^2 \\
&= 2 \sum_{j=1}^K \sum_{i=1}^m \left( \int_{(j+i-2)/n}^{(j+i-1)/n} \sigma_u^2 du \right)^2 + \sum_{j=1}^K \sum_{i=1}^m \sum_{i'=1}^m \sigma_{j,i}^2 \sigma_{j,i'}^2 \\
&= O \left( \frac{m^2}{n} \right).
\end{aligned}$$

■

## A.2 Proof of Proposition 2

### A.2.1 Notation and preliminary calculations

Let

$$z_{j,i} = \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u,$$

and so

$$\widehat{\theta}_{n,K,j} = \sum_{i=1}^m z_{j,i}^2,$$

where:  $z_{j,i}^2 = \sigma_{j,i}^2 + y_{j,i}$ ,  $\sigma_{j,i}^2 = E[(\int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u)^2] = \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du$ , and  $y_{j,i} = (\int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u)^2 - \sigma_{j,i}^2$ , which are independent (across  $i$  for given  $j$ ) and mean zero random variables conditional on the process  $\sigma_u^2$ . That is,  $E[z_{j,i} z_{j,i'}] = E[y_{j,i} y_{j,i'}] = 0$  whenever  $i \neq i'$ . Furthermore,  $z_{j,i}$  is conditionally normal and so satisfies  $E[z_{j,i}^4] = 3E^2[z_{j,i}^2]$ . Note that, for  $j < j'$ ,

$$E[z_{j,i} z_{j',i'}] = \begin{cases} \int_{(\max\{j,j'\}+(i-1)K)/n}^{(\min\{j,j'\}+iK)/n} \sigma_u^2 du \neq 0 & i = i' \\ \int_{(j+i'K)/n}^{(j'+i'K)/n} \sigma_u^2 du \neq 0 & i = i' + 1 \\ 0 & |i - i'| > 1, i = i' - 1. \end{cases}$$

Furthermore,  $E[y_{j,i}z_{j,i}] = 0$ , and

$$\begin{aligned}\text{var}(z_{j,i}) &= E \left[ \left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u \right)^2 \right] = \sigma_{j,i}^2 = \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du = O_p(K/n) \\ \text{var}(y_{j,i}) &= E \left[ \left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u \right)^4 \right] - E^2 \left[ \left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u dW_u \right)^2 \right] \\ &= 2 \left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du \right)^2 \\ &= \frac{2K}{n} \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^4 du + o_p(K^2/n^2) = O_p(K^2/n^2),\end{aligned}$$

since

$$\left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du \right)^2 = \frac{K}{n} \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^4 du + o(K^2/n^2). \quad (11)$$

We prove this result below.

By adding these results over all subsamples, we get that for each  $j$  and each  $j'$ ,

$$\text{Cov}(\hat{\theta}_{n,m,j}, \hat{\theta}_{n,m,j'}) = O(m^{-1}). \quad (12)$$

PROOF OF (11). This is established as follows. Suppose that  $f$  is a bounded positive continuous function on  $[0, 1]$ , say. Then for  $\epsilon > 0$ ,  $g(\epsilon) = \int_0^\epsilon f(x) dx$  is of order  $\epsilon$  as  $\epsilon \rightarrow 0$ , because

$$\int_0^\epsilon f(x) dx \leq \epsilon \times \sup_{x \in [0, \epsilon]} f(x).$$

Furthermore,  $g$  is differentiable in  $\epsilon$  with  $g'(\epsilon) = f(\epsilon)$ . Therefore, by the mean value theorem  $g(\epsilon) = g(0) + \epsilon g'(\bar{\epsilon}) = \epsilon f(\bar{\epsilon})$  for some  $\bar{\epsilon} \leq \epsilon$ . Therefore,

$$\left( \int_0^\epsilon f(x) dx \right)^2 = \epsilon^2 f^2(\bar{\epsilon}).$$

Furthermore,  $f^2$  is also a bounded continuous function on  $[0, 1]$  and satisfies

$$\int_0^\epsilon f^2(x) dx \leq \epsilon \times \sup_{x \in [0, \epsilon]} f^2(x)$$

$$\int_0^\epsilon f^2(x)dx = \epsilon f^2(\epsilon^*)$$

for some  $\epsilon^* \leq \epsilon$ . By continuity of  $f^2$  at 0,

$$\lim_{\epsilon \downarrow 0} \frac{f^2(\bar{\epsilon})}{f^2(\epsilon^*)} = 1,$$

and it follows that

$$\left( \int_0^\epsilon f(x)dx \right)^2 = \epsilon \int_0^\epsilon f^2(x)dx + o(\epsilon^2).$$

■

## A.2.2 Main proof

Write

$$\begin{aligned} \widehat{V} &= \frac{m}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j}^2 + m\widehat{\theta}_n^2 - 2m\widehat{\theta}_n \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} \\ &= \frac{m}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j}^2 + m\theta^2 - 2m\theta \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} + o_p(1), \end{aligned}$$

where the approximation is valid by Barndorff-Nielsen and Shephard (2002).

We make use of the following lemmas.

LEMMA 2.1.

$$\mathbb{E} \left[ \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} \right] = \frac{1}{K} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du.$$

LEMMA 2.2.

$$\mathbb{E} \left[ \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j}^2 \right] = \frac{2}{n} \sum_{j=1}^K \int_{j/n}^{1-j/n} \sigma_u^4 du + \frac{1}{K} \sum_{j=1}^K \left( \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du \right)^2 + o\left(\frac{1}{m}\right).$$

LEMMA 2.3.

$$\text{Var} \left( \frac{m}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j}^2 - 2m\theta \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} \right) = O(1).$$

These lemmas imply that

$$\begin{aligned}
\mathbb{E}\widehat{V} &= \frac{m}{K} \sum_{j=1}^K \mathbb{E}[\widehat{\theta}_{n,m,j}^2] + m\theta^2 - 2m\theta \frac{1}{K} \sum_{j=1}^K \mathbb{E}[\widehat{\theta}_{n,m,j}] + o(1) \\
&= \frac{2m}{n} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^4 du + \frac{m}{K} \sum_{j=1}^K \left( \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du \right)^2 + m \left( \int_0^1 \sigma_u^2 du \right)^2 \\
&\quad - 2m \left( \int_0^1 \sigma_u^2 du \right) \frac{1}{K} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du + o(1).
\end{aligned}$$

We have to show that

$$R = \left( \int_0^1 \sigma_u^2 du \right)^2 + \frac{1}{K} \sum_{j=1}^K \left\{ \left( \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du \right)^2 - 2 \int_0^1 \sigma_u^2 du \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du \right\} = o(K/n). \tag{13}$$

This is true because, write  $I = \int_0^1 \sigma_u^2 du$ ,  $I_{j/n} = \int_0^{j/n} \sigma_u^2 du$ , and  $I^{1-K/n+j/n} = \int_{1-K/n+j/n}^1 \sigma_u^2 du$ .

Then

$$\begin{aligned}
R &= I^2 + \frac{1}{K} \sum_{j=1}^K \{ I^2 - 2I(I_{j/n} + I^{1-K/n+j/n}) + (I_{j/n} + I^{1-K/n+j/n})^2 - 2I(I - (I_{j/n} + I^{1-K/n+j/n})) \} \\
&= \frac{1}{K} \sum_{j=1}^K (I_{j/n} + I^{1-K/n+j/n})^2 = O(K^2/n^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\widehat{V} &= \frac{2m}{n} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^4 du + o(1) \\
&= \frac{2m}{n} \sum_{j=1}^K \int_0^1 \sigma_u^4 du + o(1) \\
&= 2 \int_0^1 \sigma_u^4 du + o(1),
\end{aligned}$$

and the first part of (6) follows. The second part follows from Lemma 2.3. ■

PROOF OF LEMMA 2.1. We have

$$\begin{aligned}
E \left[ \frac{1}{K} \sum_{j=1}^K \widehat{\theta}_{n,m,j} \right] &= \frac{1}{K} \sum_{j=1}^K \sum_{i=1}^m E [z_{j,i}^2] \\
&= \frac{1}{K} \sum_{j=1}^K \sum_{i=1}^m \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du \\
&= \frac{1}{K} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du.
\end{aligned}$$

■

PROOF OF LEMMA 2.2. We have

$$\begin{aligned}
\frac{1}{K} \sum_{j=1}^K E \left( \widehat{\theta}_{n,m,j}^2 \right) &= \frac{1}{K} \sum_{j=1}^K E \left[ \left( \sum_{i=1}^m z_{j,i}^2 \right)^2 \right] \\
&= \frac{1}{K} \sum_{j=1}^K E \left[ \left( \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) \right)^2 \right] \\
&= \frac{1}{K} \sum_{j=1}^K \sum_{i=1}^m E y_{j,i}^2 + \frac{1}{K} \sum_{j=1}^K \sum_{i=1}^m \sum_{i'=1}^m \sigma_{j,i}^2 \sigma_{j,i'}^2 \\
&= \frac{2}{n} \sum_{j=1}^K \sum_{i=1}^m \left( \int_{(j+(i-1)K)/n}^{(j+iK)/n} \sigma_u^2 du \right)^2 + \frac{1}{K} \sum_{j=1}^K \left( \int_{j/n}^{1-j/n} \sigma_u^2 du \right)^2 \\
&= \frac{2}{n} \sum_{j=1}^K \int_{j/n}^{1-K/n+j/n} \sigma_u^4 du + \frac{1}{K} \sum_{j=1}^K \left( \int_{j/n}^{1-K/n+j/n} \sigma_u^2 du \right)^2 + o \left( \frac{1}{m} \right).
\end{aligned}$$

by (11).

■

PROOF OF LEMMA 2.3. This can be seen from the covariance between  $\widehat{\theta}_{n,K,j}$  and  $\widehat{\theta}_{n,K,i}$ .

■

### A.3 Proof of Proposition 3

We first derive the mean and variance of  $\widehat{\theta}_{n,m,j}$ . We have

$$\begin{aligned}
\mathbb{E}\widehat{\theta}_{n,m,j} &= K\mathbb{E}\sum_{i=1}^m (X_{j+(i-1)K} - X_{j+(i-1)K-1})^2 \\
&= K\sum_{i=1}^m \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u^2 du \\
&= \int_0^1 \sigma_u^2 du + o(1)
\end{aligned}$$

by Riemann integrability of  $\sigma_u^2$ . Furthermore,

$$\begin{aligned}
\text{Var}\widehat{\theta}_{n,m,j} &= K^2\text{Var}\sum_{i=1}^m (X_{j+(i-1)K} - X_{j+(i-1)K-1})^2 \\
&= K^2\sum_{i=1}^m \text{Var}(X_{j+(i-1)K} - X_{j+(i-1)K-1})^2 \\
&= 2K^2\sum_{i=1}^m \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u^2 du \right)^2 \\
&= \frac{2}{m} \int_0^1 \sigma_u^4 du + o\left(\frac{1}{m}\right).
\end{aligned}$$

Now we calculate the expected value of the estimator.

$$\begin{aligned}
\mathbb{E}\widehat{V}_{\text{Infill returns}} &= \mathbb{E}m \times \frac{1}{K} \sum_{j=1}^K (\widehat{\theta}_{n,m,j} - \widehat{\theta}_n)^2 \\
&= \frac{m}{K} \sum_{j=1}^K \mathbb{E}(\widehat{\theta}_{n,m,j} - \mathbb{E}\widehat{\theta}_{n,m,j})^2 + R \\
&= 2 \int_0^1 \sigma_u^4 du + o(1) + R,
\end{aligned}$$

where

$$\begin{aligned}
R &= \frac{m}{K} \sum_{j=1}^K \mathbb{E} \left\{ (\widehat{\theta}_{n,m,j} - \mathbb{E}\widehat{\theta}_{n,m,j}) (\mathbb{E}\widehat{\theta}_{n,m,j} - \widehat{\theta}_n) \right\} + \frac{m}{K} \sum_{j=1}^K \mathbb{E} \left\{ (\mathbb{E}\widehat{\theta}_{n,m,j} - \widehat{\theta}_n)^2 \right\} \\
&= \frac{m}{K} \sum_{j=1}^K \mathbb{E} \left\{ O_p(m^{-1/2}) (\mathbb{E}\widehat{\theta}_{n,m,j} - \widehat{\theta}_n) \right\} + \frac{m}{K} \sum_{j=1}^K \mathbb{E} \left\{ (\mathbb{E}\widehat{\theta}_{n,m,j} - \widehat{\theta}_n)^2 \right\}.
\end{aligned}$$

We know that  $\widehat{\theta}_n - \theta = O(n^{-1/2})$ . Therefore, have  $R = o(1)$  if  $E\widehat{\theta}_{n,m,j} - \theta = o_p(m^{-1/2})$ . For this, Riemann integrability is not enough, which is why we have assumed Hölder continuity of order larger than  $1/2$ .

Now we calculate the variance of  $\widehat{V}_{\text{Infill returns}}$ . To facilitate calculations, introduce the usual notation:

$$\begin{aligned} z_{j,i} &= \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \\ \widehat{\theta}_{n,m,j} &= K \sum_{i=1}^m z_{j,i}^2 \\ z_{j,i}^2 &= \sigma_{j,i}^2 + y_{j,i} \end{aligned}$$

$$\begin{aligned} \sigma_{j,i}^2 &= E\left[\left(\int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u\right)^2\right] = \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u^2 du \\ y_{j,i} &= \left(\int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u\right)^2 - \sigma_{j,i}^2. \end{aligned}$$

We first do some preliminary calculations.

$$\begin{aligned} Ey_{j,i} &= 0, \text{ so that } \text{Var}(y_{j,i}) = Ey_{j,i}^2 \\ Ey_{j,i}^2 &= E\left\{\left(\int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u\right)^2 - \sigma_{j,i}^2\right\}^2 \\ &= E\left\{\left(\int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u\right)^4 + (\sigma_{j,i}^2)^2 - 2\left(\int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u\right)^2 \sigma_{j,i}^2\right\} \\ &= 3(\sigma_{j,i}^2)^2 + (\sigma_{j,i}^2)^2 - 2(\sigma_{j,i}^2)^2 = 2(\sigma_{j,i}^2)^2 \end{aligned}$$

$$\begin{aligned}
\mathbb{E}y_{j,i}^4 &= \mathbb{E} \left\{ \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \right)^2 - \sigma_{j,i}^2 \right\}^4 \\
&= \mathbb{E} \left\{ \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \right)^8 + (\sigma_{j,i}^2)^4 - 4 \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \right)^2 (\sigma_{j,i}^2)^3 \right. \\
&\quad \left. - 4 \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \right)^6 (\sigma_{j,i}^2) + 6 \left( \int_{(j+(i-1)K-1)/n}^{(j+(i-1)K)/n} \sigma_u dW_u \right)^4 (\sigma_{j,i}^2)^2 \right\} \\
&= 105 (\sigma_{j,i}^2)^4 + (\sigma_{j,i}^2)^4 - 4 (\sigma_{j,i}^2) (\sigma_{j,i}^2)^3 - 4 \times 15 (\sigma_{j,i}^2)^3 (\sigma_{j,i}^2) + 6 \times 3 (\sigma_{j,i}^2)^2 (\sigma_{j,i}^2)^2 \\
&= (\sigma_{j,i}^2)^4 \{105 + 1 - 4 - 60 + 18\} = 60 (\sigma_{j,i}^2)^4.
\end{aligned}$$

Using the same approximation of  $\widehat{V}_{\text{Infill returns}}$  as in Proposition 1, the variance of  $\widehat{V}_{\text{Infill returns}}$  becomes

$$\begin{aligned}
\text{Var} \widehat{V}_{\text{Infill returns}} &= \text{Var} \left[ \frac{m}{K} \sum_{j=1}^K (\widehat{\theta}_{n,m,j} - \theta)^2 + o_p(1) \right] \\
&= \frac{m^2}{K^2} \sum_{j=1}^K \text{Var} (\widehat{\theta}_{n,m,j} - \theta)^2 + o(1).
\end{aligned}$$

To show  $\text{Var} \widehat{V}_{\text{Infill returns}} = o(1)$ , we show that  $\text{Var} (\widehat{\theta}_{n,m,j} - \theta)^2 = \text{Var} (\widehat{\theta}_{n,m,j}^2 - 2\widehat{\theta}_{n,m,j}\theta) = o(Km^{-2})$ . In order to do that, we proceed in three steps. That is to say, to calculate  $\text{Var}(x)$ , we first first calculate  $E(x)$ , then  $y = x - E(x)$ , and finally  $E(y^2)$ .

Step 1. We have

$$\begin{aligned}
&\mathbb{E} (\widehat{\theta}_{n,m,j}^2 - 2\widehat{\theta}_{n,m,j}\theta) \\
&= \mathbb{E} \left[ \left\{ K \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) \right\}^2 - 2\theta K \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) \right] \\
&= \mathbb{E} \left\{ K^2 \left( \sum_{i=1}^m \sigma_{j,i}^2 \right)^2 + K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 + 2K^2 \sum_{i=1}^m \sigma_{j,i}^2 \sum_{i=1}^m y_{j,i} \right\} - 2\theta K \sum_{i=1}^m \sigma_{j,i}^2
\end{aligned}$$

$$\begin{aligned}
&= K^2 \left( \sum_{i=1}^m \sigma_{j,i}^2 \right)^2 + K^2 \sum_{i=1}^m \mathbb{E} y_{j,i}^2 - 2\theta K \sum_{i=1}^m \sigma_{j,i}^2 \\
&= K^2 \left( \sum_{i=1}^m \sigma_{j,i}^2 \right)^2 + 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 - 2\theta K \sum_{i=1}^m \sigma_{j,i}^2.
\end{aligned}$$

Step 2. We have

$$\begin{aligned}
&\widehat{\theta}_{n,m,j}^2 - 2\widehat{\theta}_{n,m,j}\theta - \mathbb{E} \left( \widehat{\theta}_{n,m,j}^2 - 2\widehat{\theta}_{n,m,j}\theta \right) \\
&= \left\{ K^2 \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) \right\}^2 - 2\theta K \sum_{i=1}^m (\sigma_{j,i}^2 + y_{j,i}) - \mathbb{E} \left( \widehat{\theta}_{n,m,j}^2 - 2\widehat{\theta}_{n,m,j}\theta \right) \\
&= K^2 \left( \sum_{i=1}^m \sigma_{j,i}^2 \right)^2 + K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 + 2K^2 \sum_{i=1}^m \sigma_{j,i}^2 \sum_{i=1}^m y_{j,i} \\
&\quad - 2\theta K \sum_{i=1}^m \sigma_{j,i}^2 - 2\theta K \sum_{i=1}^m y_{j,i} \\
&\quad - K^2 \left( \sum_{i=1}^m \sigma_{j,i}^2 \right)^2 - 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 + 2\theta K \sum_{i=1}^m \sigma_{j,i}^2 \\
&= K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 + 2K^2 \sum_{i=1}^m \sigma_{j,i}^2 \sum_{i=1}^m y_{j,i} - 2\theta K \sum_{i=1}^m y_{j,i} - 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 \\
&= K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 - 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 + 2K \sum_{i=1}^m y_{j,i} \left( K \sum_{i=1}^m \sigma_{j,i}^2 - \theta \right) \\
&= K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 - 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 + o_p(n^{-1}),
\end{aligned}$$

because even without any smoothness assumptions on  $\sigma$  we have  $K \sum_{i=1}^m \sigma_{j,i}^2 - \theta = o(1)$  and  $K \sum_{i=1}^m y_{j,i} = O(Kmn^{-2}) = O(n^{-1})$ .

Step 3. The only term in the variance, whose negligibility needs to be shown (because  $\sqrt{Km^{-2}} >$

$n^{-1}$ ), is

$$\begin{aligned}
& \mathbb{E} \left[ K^2 \left( \sum_{i=1}^m y_{j,i} \right)^2 - 2K^2 \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right]^2 \\
&= K^4 \left[ \mathbb{E} \left( \sum_{i=1}^m y_{j,i} \right)^4 + 4\mathbb{E} \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 - 4 \sum_{i=1}^m (\sigma_{j,i}^2)^2 \mathbb{E} \left( \sum_{i=1}^m y_{j,i} \right)^2 \right] \\
&= K^4 \left[ \mathbb{E} \left( \sum_{i=1}^m y_{j,i} \right)^4 + 4 \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 - 8 \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 \right] \\
&= K^4 \left[ \mathbb{E} \left( \sum_{i=1}^m y_{j,i} \right)^4 - 4 \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 \right] \\
&= K^4 \left[ \sum_{i=1}^m \mathbb{E} y_{j,i}^4 + 3 \sum_{i'=1, i' \neq i}^m \sum_{i=1}^m \mathbb{E} y_{j,i}^2 \mathbb{E} y_{j,i'}^2 - 4 \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 \right] \\
&= K^4 \left[ 60 \sum_{i=1}^m (\sigma_{j,i}^2)^4 + 12 \sum_{i'=1, i' \neq i}^m \sum_{i=1}^m (\sigma_{j,i}^2)^2 (\sigma_{j,i'}^2)^2 - 4 \left( \sum_{i=1}^m (\sigma_{j,i}^2)^2 \right)^2 \right] \\
&= K^4 \left[ 56 \sum_{i=1}^m (\sigma_{j,i}^2)^4 + 8 \sum_{i'=1, i' \neq i}^m \sum_{i=1}^m (\sigma_{j,i}^2)^2 (\sigma_{j,i'}^2)^2 \right] \\
&= O(K^4 m n^{-4}) + O(K^4 m^2 n^{-4}) = O(m^{-2}) = o(K m^{-2}).
\end{aligned}$$

This implies  $\text{Var} \widehat{V} \rightarrow 0$  and so we have mean square convergence and so (8) follows by Chebyshev's inequality.  $\blacksquare$

#### A.4 Proof of Proposition 4

In the first subsection of this proof we explain the notation, in the second we show that the bias of  $\widehat{V}_{SC}$  is negligible, i.e.,  $\mathbb{E}(\widehat{V}_{SC}) - V = o(1)$ , and in the third subsection we show that  $\text{Var}(\widehat{V}_{SC}) = o(1)$ , so that Proposition 4 follows by Chebyshev's inequality.

#### A.4.1 Notation for $\widehat{V}_{SC}$

Much of the notation is the same as for the other estimators.  $K$  is the number of subsamples.  $m$  is the number of high frequency returns in of each subsample.  $J$  is the number of high frequency returns that ‘fit into’ one low frequency return (see Figure 2 for a graphical illustration),  $1 < J < m$ . There are  $m/J$  number of low frequency in each subsample. Take  $m \div J$  (i.e.,  $m$  divisible by  $J$ ).

The proof is for a general amount of overlap between subsamples, so introduce a variable  $s$ , for ‘shift’. If subsamples are constructed as observations in a window, we move the window by  $s/n$  to get every next subsample. For example,  $s = m$  corresponds to no overlap between subsamples and  $s = 1$  corresponds to maximum possible overlap. Assume  $m \div s$ . As we show below, we have  $K = n/s - m/s + 1$ .

The ‘subsample copies’ of the  $RV$  estimator ( $\widehat{\theta}_{n,m,j}$  in (2)) will be

$$\widehat{\theta}_{n,m,J,1} = \sum_{i=1}^{m/J} (X_{Ji} - X_{J(i-1)})^2, \quad \widehat{\theta}_{n,m,J,2} = \sum_{i=1}^{m/J} (X_{Ji+s} - X_{J(i-1)+s})^2, \quad \text{etc.},$$

so that the copy corresponding to the  $k^{\text{th}}$  subsample is

$$\widehat{\theta}_{n,m,J,k} = \sum_{i=1}^{m/J} (X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)})^2.$$

The number of subsamples is the  $k$  such that

$$J \frac{m}{J} + s(k_{\max} - 1) = n \Rightarrow k_{\max} \equiv K = n/s - m/s + 1.$$

We have

$$\mathbb{E} \widehat{\theta}_{n,m,J,k} = \sum_{i=1}^{m/J} \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 du = \int_{s(k-1)/n}^{[m+s(k-1)]/n} \sigma_u^2 du.$$

The ‘subsample copies’ of the true parameter  $QV$  ( $\theta$  in (2), except that we have a different ‘copy’ for each subsample for our centering) will be

$$\begin{aligned} \widehat{\theta}_{n,m,1} &= \sum_{i=1}^m (X_i - X_{i-1})^2, \quad \widehat{\theta}_{n,m,2} = \sum_{i=1}^m (X_{i+s} - X_{i-1+s})^2, \quad \text{etc.} \\ \widehat{\theta}_{n,m,k} &= \sum_{i=1}^m (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2. \end{aligned}$$

We have

$$\mathbb{E} \widehat{\theta}_{n,m,k} = \sum_{i=1}^m \int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2 du = \int_{s(k-1)/n}^{[m+s(k-1)]/n} \sigma_u^2 du.$$

Recall that the estimator of  $V$  is

$$\widehat{V}_{SC} = \frac{n^2}{mKJ} \sum_{k=1}^K \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2.$$

#### A.4.2 Derivation of $E(\widehat{V}_{SC})$

Introduce the following notation,

$$\begin{aligned} E \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2 &= E \widehat{\theta}_{n,m,J,k}^2 + E \widehat{\theta}_{n,m,k}^2 - 2E \widehat{\theta}_{n,m,J,k} \widehat{\theta}_{n,m,k} \\ &= A_k + B_k - 2C_k, \end{aligned}$$

so that we have  $E \widehat{V}_{SC} = \frac{n^2}{mKJ} \sum_{k=1}^K (A_k + B_k - 2C_k)$ . Then,

$$\begin{aligned} A_k &= E \widehat{\theta}_{n,m,J,k}^2 = E \left[ \sum_{i=1}^{m/J} (X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)})^2 \right]^2 \\ &= \sum_{i=1}^{m/J} E (X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)})^4 \\ &\quad + 2 \sum_{i'>i}^{m/J} \sum_{i=1}^{m/J} E (X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)})^2 E (X_{Ji'+s(k-1)} - X_{J(i'-1)+s(k-1)})^2 \\ &= 2 \sum_{i=1}^{m/J} \left( \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 \right)^2 + \left( \int_{s(k-1)/n}^{[m+s(k-1)]/n} \sigma_u^2 du \right)^2. \end{aligned}$$

Similarly,

$$\begin{aligned} B_k &= E \widehat{\theta}_{n,m,k}^2 = E \left[ \sum_{i=1}^m (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2 \right]^2 \\ &= 2 \sum_{i=1}^m \left( \int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2 \right)^2 + \left( \int_{s(k-1)/n}^{[m+s(k-1)]/n} \sigma_u^2 du \right)^2. \end{aligned}$$

In the third term, we have covariances between realised volatilities on the two grids, which we

denote as corresponding summation over low frequency returns  $\sum_{i=1}^{m/J} C_{k,i}$  as follows,

$$\begin{aligned}
C_k &= \mathbb{E}\widehat{\theta}_{n,m,J,k}\widehat{\theta}_{n,m,k} \\
&= \text{Cov}\left(\widehat{\theta}_{n,m,J,k}, \widehat{\theta}_{n,m,k}\right) + \mathbb{E}\widehat{\theta}_{n,m,J,k}\mathbb{E}\widehat{\theta}_{n,m,k} \\
&= \text{Cov}\left(\sum_{i=1}^{m/J} (X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)})^2, \sum_{i=1}^m (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2\right) + \mathbb{E}\widehat{\theta}_{n,m,J,k}\mathbb{E}\widehat{\theta}_{n,m,k} \\
&= \sum_{i=1}^{m/J} C_{k,i} + \mathbb{E}\widehat{\theta}_{n,m,J,k}\mathbb{E}\widehat{\theta}_{n,m,k}.
\end{aligned}$$

We then notice that  $C_{k,1}$  is the variance of the realised volatility over the 1<sup>st</sup> low frequency return, and

$$\begin{aligned}
C_{k,1} &= \text{Cov}\left[(X_{J+s(k-1)} - X_{s(k-1)})^2, \sum_{i=1}^m (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2\right] \\
&= \text{Cov}\left[(X_{J+s(k-1)} - X_{s(k-1)})^2, \sum_{i=1}^J (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2\right] \\
&= \text{Cov}\left[\sum_{j=1}^J (X_{j+s(k-1)} - X_{j-1+s(k-1)})^2\right. \\
&\quad \left.+ 2 \sum_{j'>j}^J \sum_{j=1}^J (X_{j+s(k-1)} - X_{j-1+s(k-1)}) (X_{j'+s(k-1)} - X_{j'-1+s(k-1)}),\right. \\
&\quad \left.\sum_{i=1}^J (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2\right] \\
&= \text{Var}\left[\sum_{i=1}^J (X_{i+s(k-1)} - X_{i-1+s(k-1)})^2\right] \\
&= 2 \sum_{i=1}^J \left(\int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2\right)^2.
\end{aligned}$$

Similarly with other  $C'_{k,i}$ s and so we have

$$\begin{aligned}
C_k &= 2 \sum_{i=1}^J \left(\int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2\right)^2 + \mathbb{E}\widehat{\theta}_{n,m,J,k}\mathbb{E}\widehat{\theta}_{n,m,k} \\
&= 2 \sum_{i=1}^J \left(\int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2\right)^2 + \left(\int_{s(k-1)/n}^{[m+s(k-1)]/n} \sigma_u^2 du\right)^2 \\
&= B_k.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathbb{E}\widehat{V}_{SC} \\
&= \frac{n^2}{mKJ} \sum_{k=1}^K \mathbb{E} \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2 = \frac{n^2}{mKJ} \sum_{k=1}^K (A_k - B_k) \\
&= \frac{n^2}{mKJ} \sum_{k=1}^K \left\{ 2 \sum_{i=1}^{m/J} \left( \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 \right)^2 - 2 \sum_{i=1}^J \left( \int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2 \right)^2 \right\} \\
&= 2 \frac{n^2}{mKJ} \sum_{k=1}^K \sum_{i=1}^{m/J} \left( \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 \right)^2 - 2 \frac{n^2}{mKJ} \sum_{k=1}^K \sum_{i=1}^m \left( \int_{[i-1+s(k-1)]/n}^{[i+s(k-1)]/n} \sigma_u^2 \right)^2.
\end{aligned}$$

We show just below that the first term converges to  $V$ . Notice that second term is like  $J^{-1}V$ , which means that one could do an easy finite sample bias correction by estimating  $V$  by  $\frac{J}{J-1}\widehat{V}_{SC}$  instead of  $\widehat{V}_{SC}$ . Note that in our simulation setup we have  $J = 15$ , so the adjustment factor is non-negligible.

We deal with summations in the first term by re-grouping the terms so that each group ‘covers’ the interval  $[0, 1]$  apart from end-effects,

$$\begin{aligned}
& \sum_{k=1}^K \sum_{i=1}^{m/J} \left( \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 \right)^2 \\
&= \sum_{p=1}^{m/s} \sum_{i=1}^{(n-m)/J} \left( \int_{J(i-1)/n+s(p-1)/n}^{Ji/n+s(p-1)/n} \sigma_u^2 \right)^2 \\
&= \sum_{p=1}^{m/s} \left[ \frac{J}{n} \int_0^1 \sigma_u^2 du + o\left(\frac{J}{n}\right) \right] \\
&= \frac{mJ}{sn} \int_0^1 \sigma_u^2 du + o\left(\frac{mJ}{sn}\right).
\end{aligned}$$

Now we can conclude asymptotic unbiasedness,

$$\begin{aligned}
\mathbb{E}\widehat{V}_{SC} &= 2 \frac{n^2}{mKJ} \left[ \frac{mJ}{sn} \int_0^1 \sigma_u^2 du + o\left(\frac{mJ}{sn}\right) \right] + O(J^{-1}) \\
&= V + o(1)
\end{aligned}$$

by using the fact that  $K = n/s - m/s + 1 = n/s + o(n/s)$ . ■

### A.4.3 Derivation of $\text{Var}(\widehat{V}_{SC})$

$$\begin{aligned}
\text{Var}\widehat{V}_{SC} &= \text{Var} \left[ \frac{n^2}{mKJ} \sum_{k=1}^K \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2 \right] \\
&= \left( \frac{n^2}{mKJ} \right)^2 \sum_{k'=1}^K \sum_{k=1}^K \text{Cov} \left[ \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2, \left( \widehat{\theta}_{n,m,J,k'} - \widehat{\theta}_{n,m,k'} \right)^2 \right] \\
&\leq \left( \frac{n^2}{mKJ} \right)^2 \sum_{k'=\max(1,k-m/s)}^{\min(K,k+m/s)} \sum_{k=1}^K \text{Cov} \left[ \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2, \left( \widehat{\theta}_{n,m,J,k'} - \widehat{\theta}_{n,m,k'} \right)^2 \right] \\
&\leq \left( \frac{n^2}{mKJ} \right)^2 2 \frac{m}{s} \sum_{k=1}^K \text{Var} \left[ \left( \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right)^2 \right],
\end{aligned}$$

where we use the fact that: 1) all these covariances must be nonnegative, 2) for a fixed term  $k$  in the first summation, only the terms from  $k - m/s$  to  $k + m/s$  terms in the summation over  $k'$  give rise to nonzero covariances.

Now we calculate the magnitude of  $\widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k}$ ,

$$\begin{aligned}
&\text{Var} \left[ \widehat{\theta}_{n,m,J,k} - \widehat{\theta}_{n,m,k} \right] \\
&= \text{Var} \left[ \sum_{i=1}^{m/J} \left( X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)} \right)^2 - \sum_{i=1}^m \left( X_{i+s(k-1)} - X_{i-1+s(k-1)} \right)^2 \right] \\
&= \text{Var} \left[ \sum_{i=1}^{m/J} \left( X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)} \right)^2 - \sum_{i=1}^{m/J} \sum_{j=1}^J \left( X_{j+J(i-1)+s(k-1)} - X_{j-1+J(i-1)+s(k-1)} \right)^2 \right] \\
&= \text{Var} \sum_{i=1}^{m/J} \left[ \left( X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)} \right)^2 - \sum_{j=1}^J \left( X_{j+J(i-1)+s(k-1)} - X_{j-1+J(i-1)+s(k-1)} \right)^2 \right] \\
&= \sum_{i=1}^{m/J} \text{Var} \left[ \left( X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)} \right)^2 - \sum_{j=1}^J \left( X_{j+J(i-1)+s(k-1)} - X_{j-1+J(i-1)+s(k-1)} \right)^2 \right] \\
&= \sum_{i=1}^{m/J} \left\{ \text{Var} \left( X_{Ji+s(k-1)} - X_{J(i-1)+s(k-1)} \right)^2 \right. \\
&\quad \left. - \sum_{j=1}^J \text{Var} \left( X_{j+J(i-1)+s(k-1)} - X_{j-1+J(i-1)+s(k-1)} \right)^2 \right\} \tag{14}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{m/J} \left\{ 2 \left( \int_{[J(i-1)+s(k-1)]/n}^{[Ji+s(k-1)]/n} \sigma_u^2 du \right)^2 - 2 \sum_{j=1}^J \left( \int_{[j-1+J(i-1)+s(k-1)]/n}^{[j+J(i-1)+s(k-1)]/n} \sigma_u^2 du \right)^2 \right\} \\
&= O \left( \frac{m}{J} \left( \frac{J}{n} \right)^2 \right) = O \left( \frac{mJ}{n^2} \right),
\end{aligned}$$

where (14) follows by noticing that, in the expression of variance, the covariance between the first and second term equals the variance of the second term.

Using this, we can show that  $\text{Var}(\widehat{V}_{SC})$  is negligible. We have

$$\text{Var}(\widehat{V}_{SC}) \sim \left( \frac{n^2}{mKJ} \right)^2 \frac{m}{s} K \left( \frac{mJ}{n^2} \right)^2 = \frac{n^4 m K m^2 J^2}{m^2 K^2 J^2 s n^4} = \frac{m}{Ks} \sim \frac{m}{n},$$

and recall we have  $m/n = o(1)$ .

Notice how the magnitude of  $\text{Var}(\widehat{V}_{SC})$  does not depend on the amount of overlap (i.e., it does not depend on the parameter  $s$ ). ■

## References

- [1] AÏT-SAHALIA, Y., AND J. JACOD (2006). Testing for jumps in a discretely observed process. Unpublished paper: Department of Economics, Princeton University.
- [2] AÏT-SAHALIA, Y., P. MYKLAND, AND L. ZHANG (2005). How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Review of Financial Studies*, 18, 351-416.
- [3] AÏT-SAHALIA, Y., P. MYKLAND, AND L. ZHANG (2006a). Ultra high frequency volatility estimation with dependent microstructure noise. Unpublished paper: Department of economics, Princeton University
- [4] AÏT-SAHALIA, Y., P. MYKLAND, AND L. ZHANG (2006b). Comments on ‘Realized Variance and Market Microstructure Noise,’ by P. Hansen and A. Lunde, *Journal of Business and Economic Statistics*.
- [5] AÏT-SAHALIA, Y., ZHANG, L. , AND P. MYKLAND (2005). Edgeworth Expansions for Realized Volatility and Related Estimators. Working paper, Princeton University.
- [6] AWARTANI, B., CORRADI, V., AND W. DISTASO (2004). Testing and Modelling Market Microstructure Effects with and Application to the Dow Jones Industrial Average. Working paper.
- [7] BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A., AND N. SHEPHARD (2006). Designing Realised Kernels to Measure the Ex-post Variation of Equity Prices in the Presence of Noise. Working Paper.
- [8] BARNDORFF-NIELSEN, O. E. AND SHEPHARD, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B* 64, 253–280.
- [9] BARNDORFF-NIELSEN, O. E. AND SHEPHARD, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* 4, 1-30.
- [10] BARNDORFF-NIELSEN, O. E. AND SHEPHARD, N. (2007). Variation, jumps, market frictions and high frequency data in financial econometrics. *Advances in Economics and Econometrics*.

Theory and Applications, Ninth World Congress, (edited by Richard Blundell, Persson Torsten and Whitney K Newey), Econometric Society Monographs, Cambridge University Press.

- [11] GONÇALVES, S. AND N. MEDDAHI (2005). Bootstrapping Realized Volatility. Unpublished paper.
- [12] HOROWITZ, J.L. (2001). The Bootstrap. in *The Handbook of Econometrics, vol 5*. Eds J.J. Heckman and E. Leamer. Reed Elsevier, Amsterdam.
- [13] HUANG, X. AND G. TAUCHEN (2005). The Relative Contribution of Jumps to Total Price Variance. *Journal of Financial Econometrics* 3(4), 456-499.
- [14] KALNINA, I. AND O. B. LINTON (2006). Estimating Quadratic Variation Consistently in the Presence of Correlated Measurement Error. STICERD Working Paper, London School of Economics
- [15] KARATZAS, I. AND S. E. SHREVE (2005). *Brownian Motion and Stochastic Calculus*, New York: Springer-Verlag.
- [16] LAHIRI, S. N., M. S. KAISER, N. CRESSIE, AND N. HSU (1999). Prediction of Spatial Cumulative Distribution Functions Using Subsampling. *Journal of the American Statistical Association*, 94, 86–97
- [17] LINTON, O.B. (2000) Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502-523.
- [18] MYKLAND, IP. AND L. ZHANG (2007). Locally parametric inference in high frequency data. University of Chicago.
- [19] PODOLSKIJ, M. (2006). New Theory on Estimation of Integrated Volatility with Applications. PhD Thesis, Bochum University
- [20] POLITIS, D. N. AND J. P. ROMANO (1994), “Large sample confidence regions based on subsamples under minimal assumptions.” *Annals of Statistics* 22, 2031-2050.
- [21] POLITIS, D. N., J. P. ROMANO AND M. WOLF (1999). *Subsampling*, Springer-Verlag, New York.

- [22] SAMORODNITSKY, G., AND M.S. TAQQU (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York.
- [23] VERAART, A. (2007). Feasible inference for realised variance in the presence of jumps. Unpublished paper, Oxford university.
- [24] ZHANG, L. (2004). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*. Forthcoming.
- [25] ZHANG, L., P. MYKLAND, AND Y. AÏT-SAHALIA (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100, 1394–1411
- [26] ZHOU, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics* 14, 45–52.

## B Figures

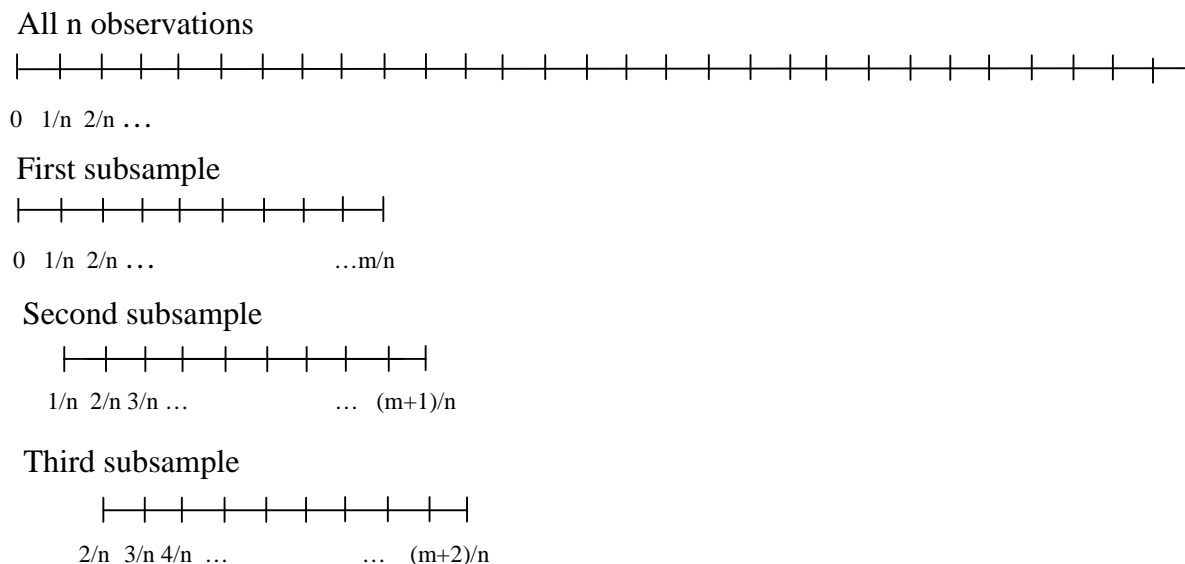


Figure 1. Regular Subsampling

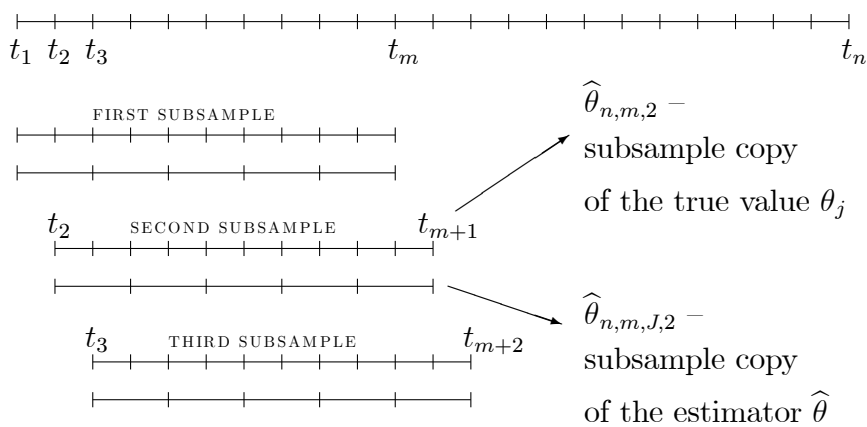


Figure 2. Subset Centered Subsampling.

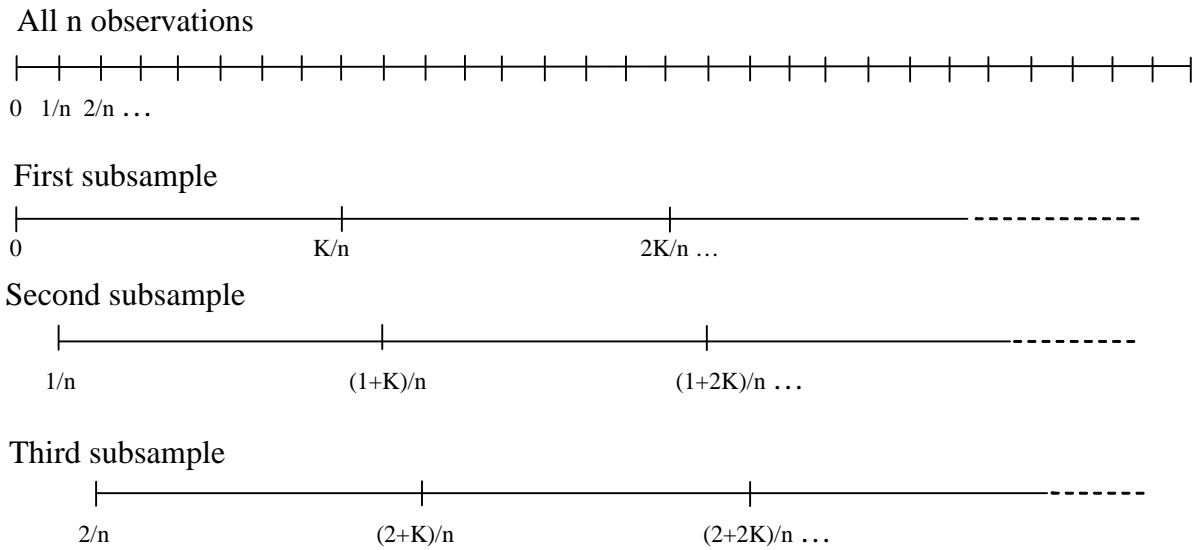


Figure 3. Infill Price Subsampling

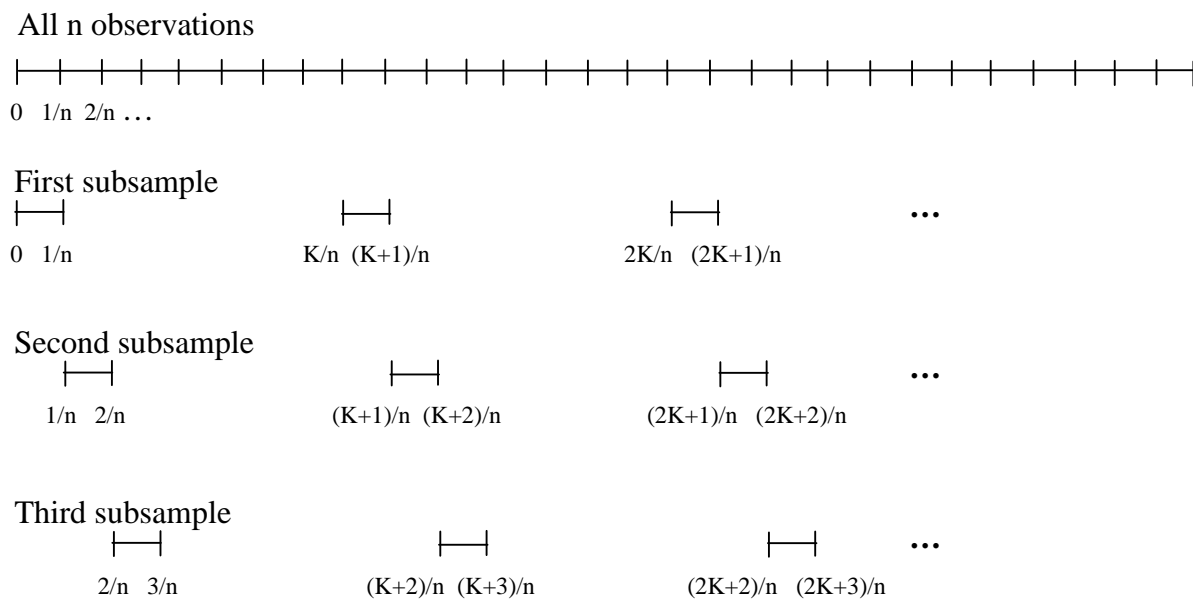


Figure 4. Infill Returns Subsampling

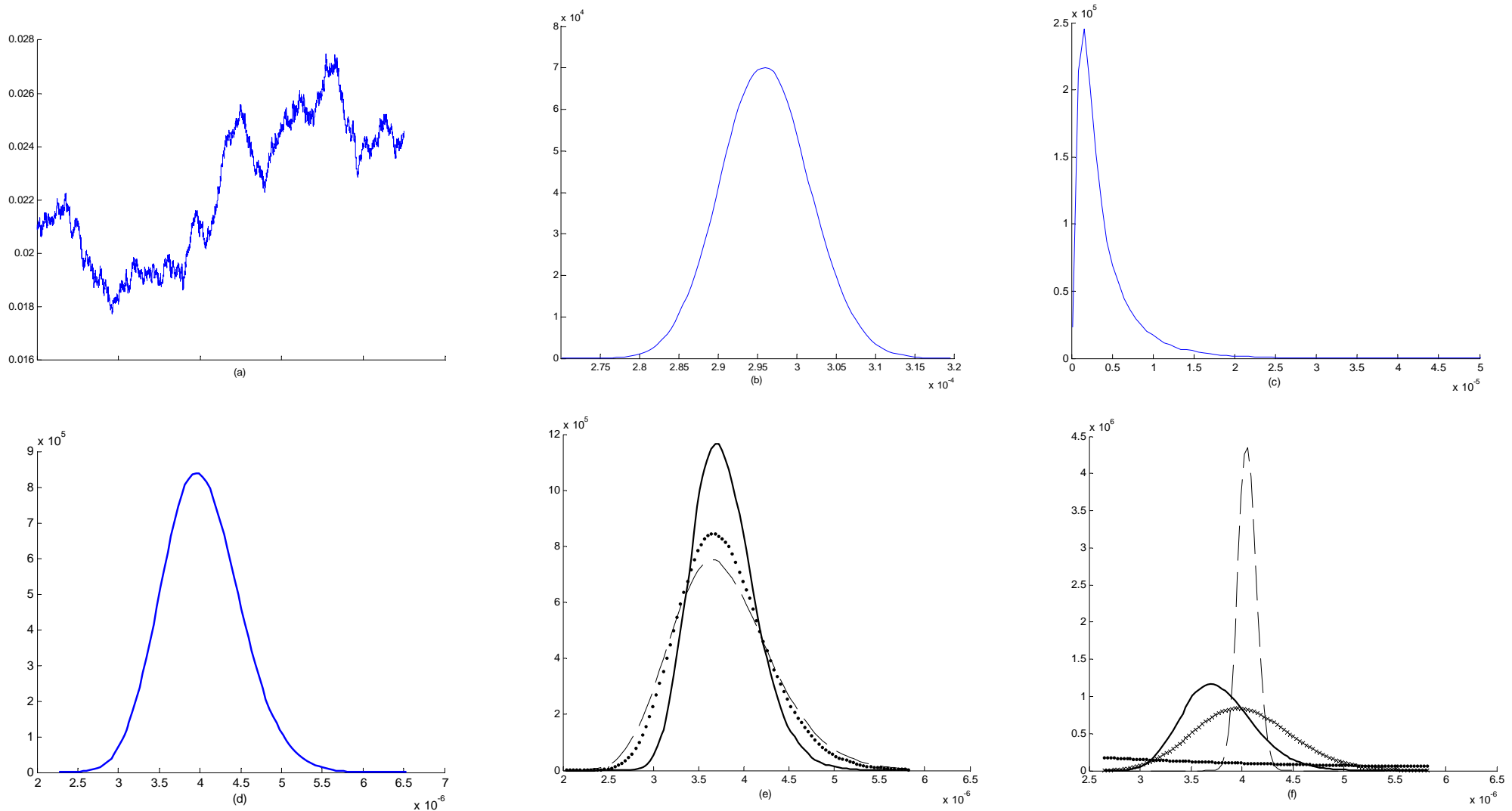
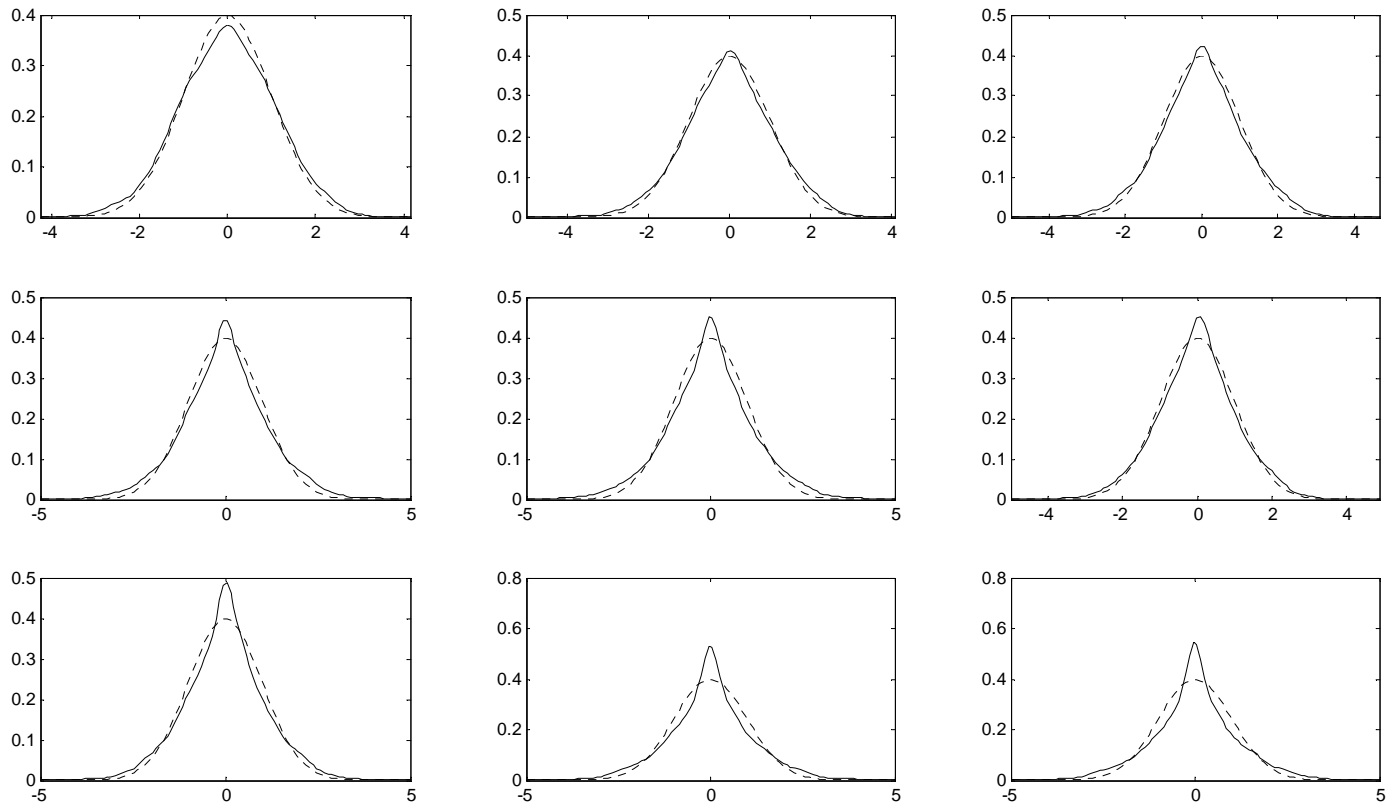
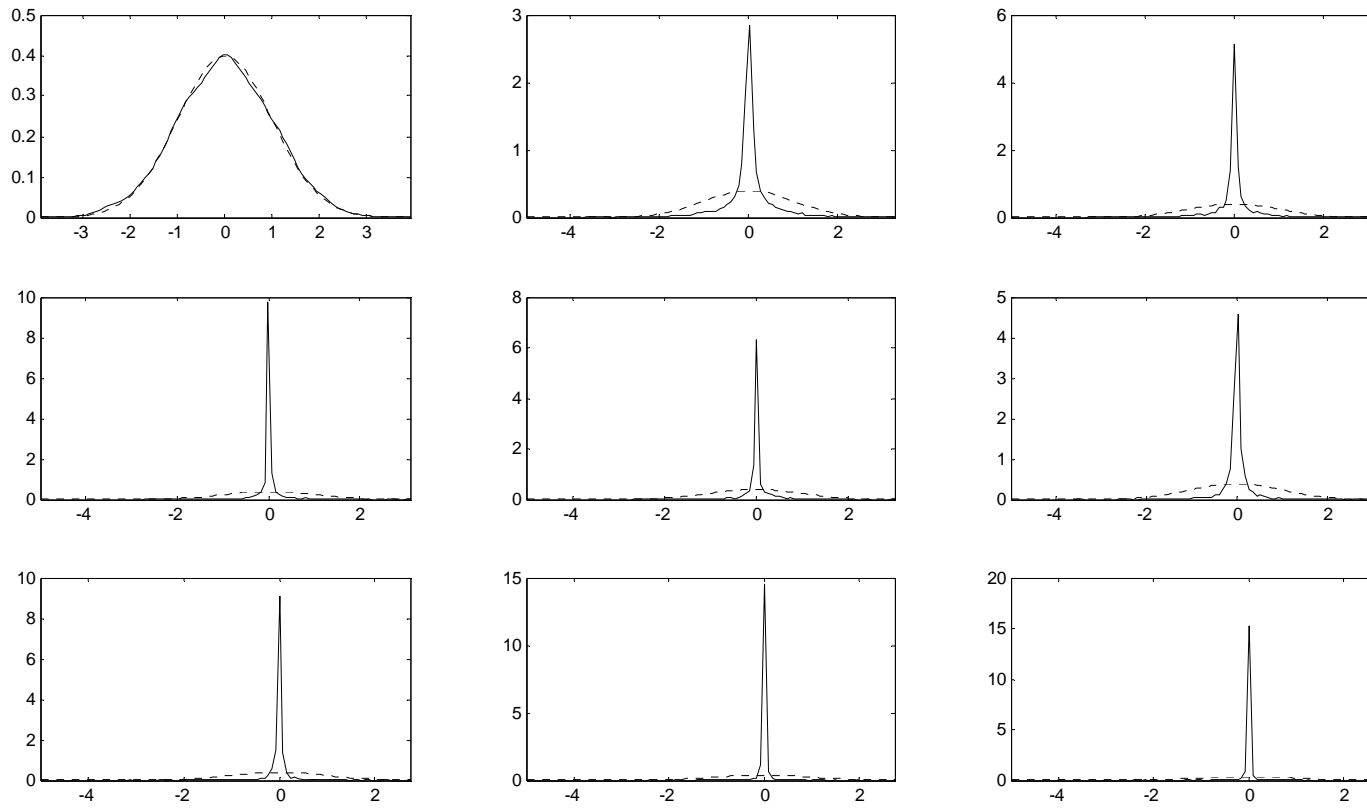


Figure (5) (a) Volatility sample path. From here, the true variance of RV is  $V = 4.05 * 10^{-6}$ .  
 (b) Kernel density over simulations of Regular Subsampling estimator. This estimator very much overestimates the true quantity.  
 (c) Kernel density over simulations of Infill Price Subsampling estimator.  
 (d) Kernel density over simulations of Infill Returns Subsampling estimator.  
 (e) Kernel densities over simulations of Subset Centered Infill estimator, for three different amounts of overlap between subsamples. Amount of overlap does not seem to affect the expected value, but it decreases the variance.  
 (f) Kernel densities over simulations of Infill Price subsampling estimator (dotted), Infill Returns Subsampling estimator (crosses), Subset Centered Infill Subsampling estimator (solid) and  $2IQ_n$  (dashed).



**Figure 6.**  
 Solid line: Estimated kernel density of studentised RV, using the Subset Centered Infill Subsampling estimator of  $V$ .  
 Dashed line: standard normal density.  
 Nine scenarios as described in Table 3.

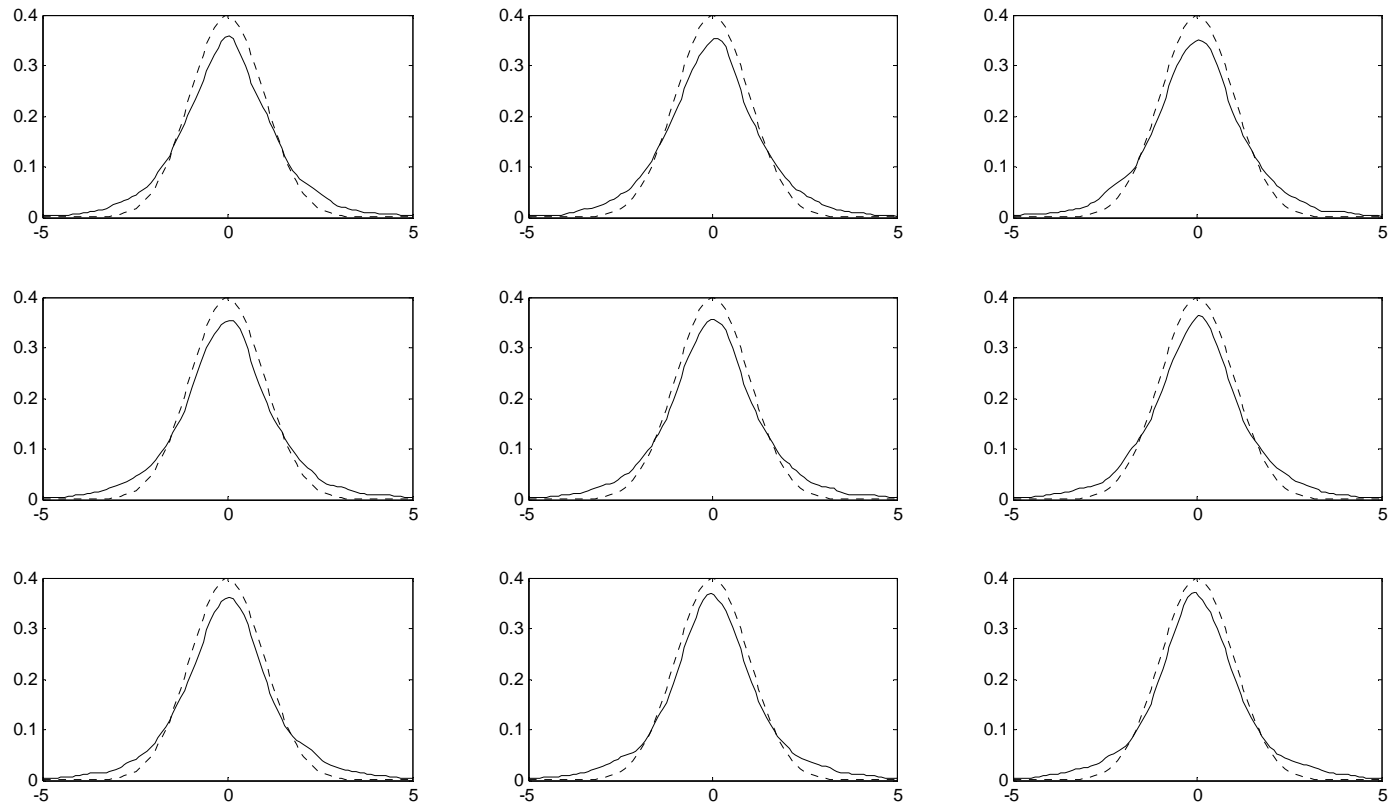


**Figure 7.**

Solid line: Estimated kernel density of studentised RV, using the Infill Returns Subsampling estimator of  $V$ .

Dashed line: standard normal density.

Nine scenarios as described in Table 3.

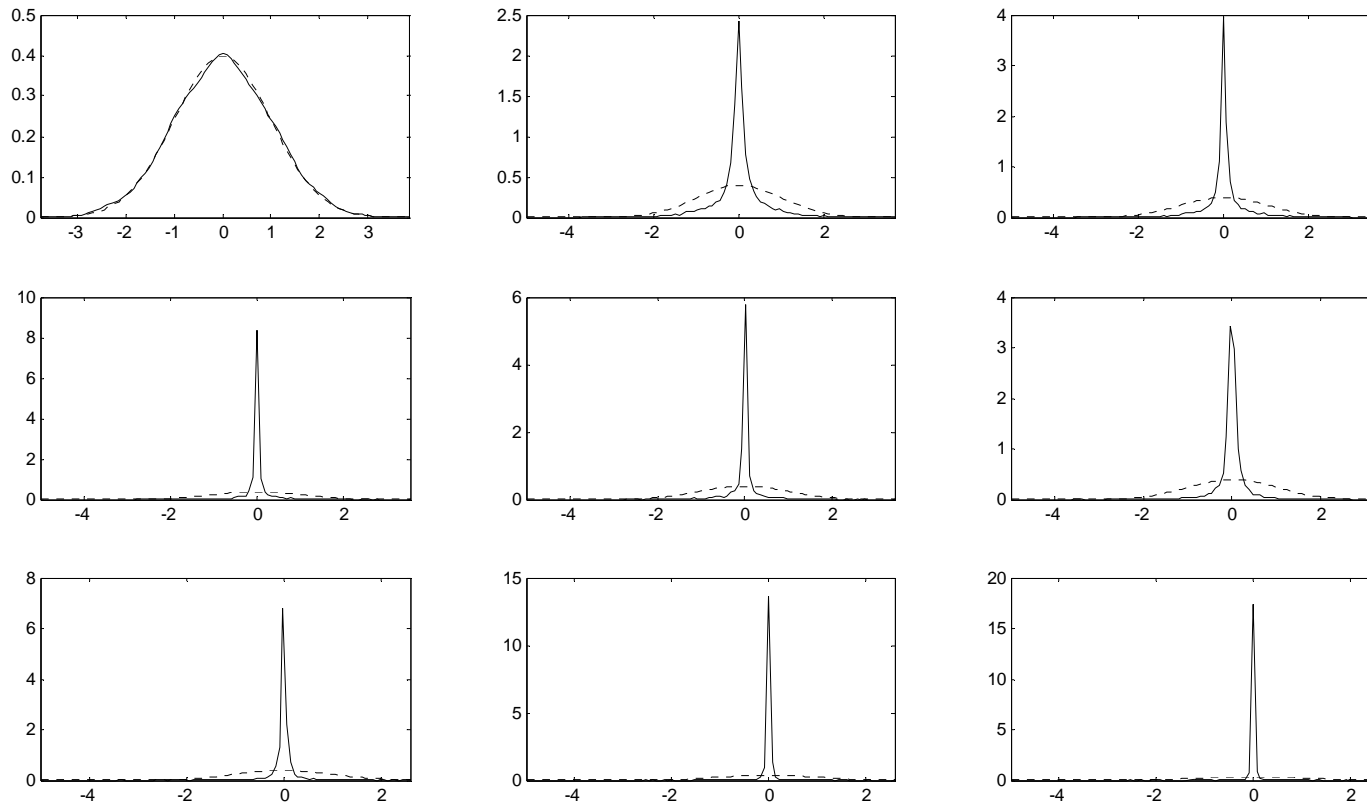


**Figure 8.**

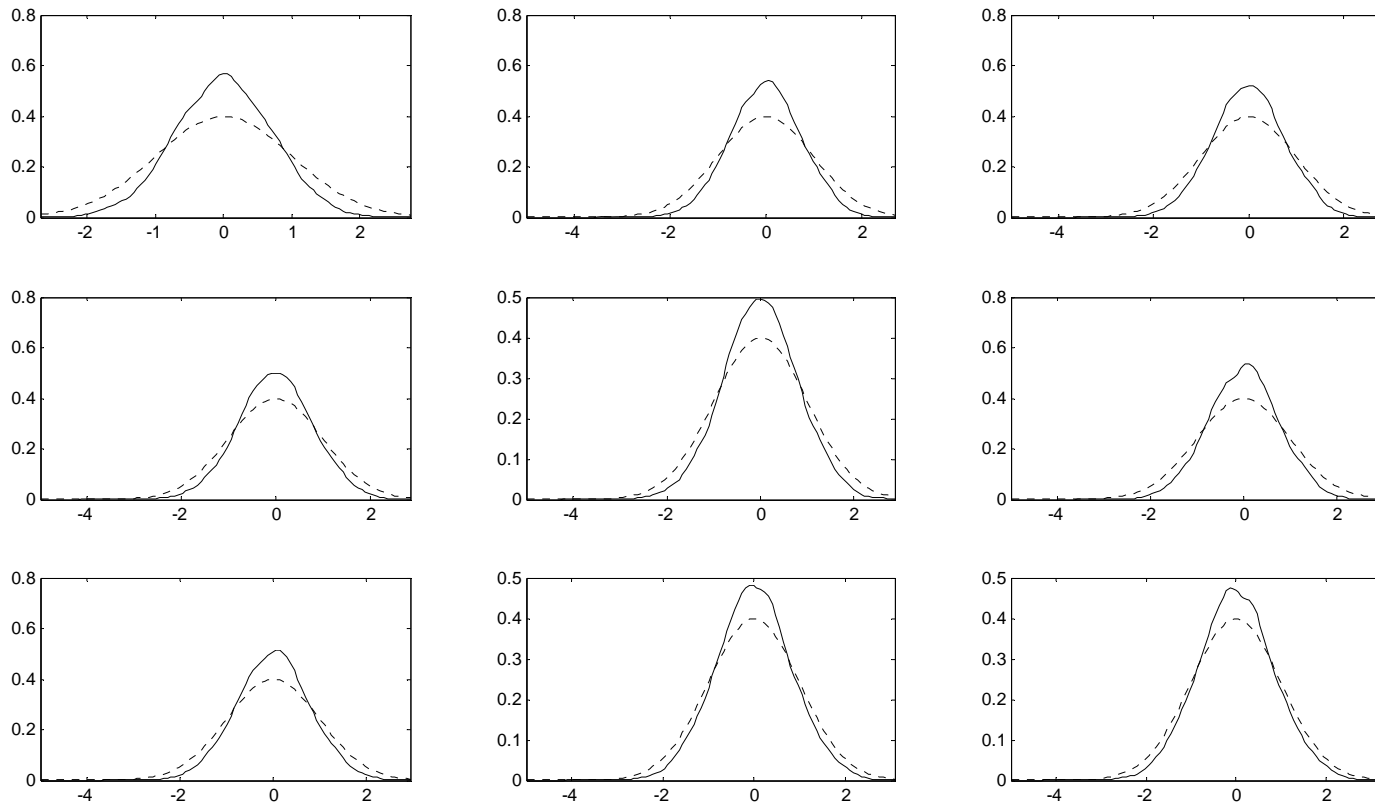
Solid line: Estimated kernel density of studentised RV, using the Infill Price Subsampling estimator of  $V$ .

Dashed line: standard normal density.

Nine scenarios as described in Table 3.



**Figure 9.**  
 Solid line: Estimated kernel density of studentised RV, using  $2IQ_n$ .  
 Dashed line: standard normal density.  
 Nine scenarios as described in Table 3.

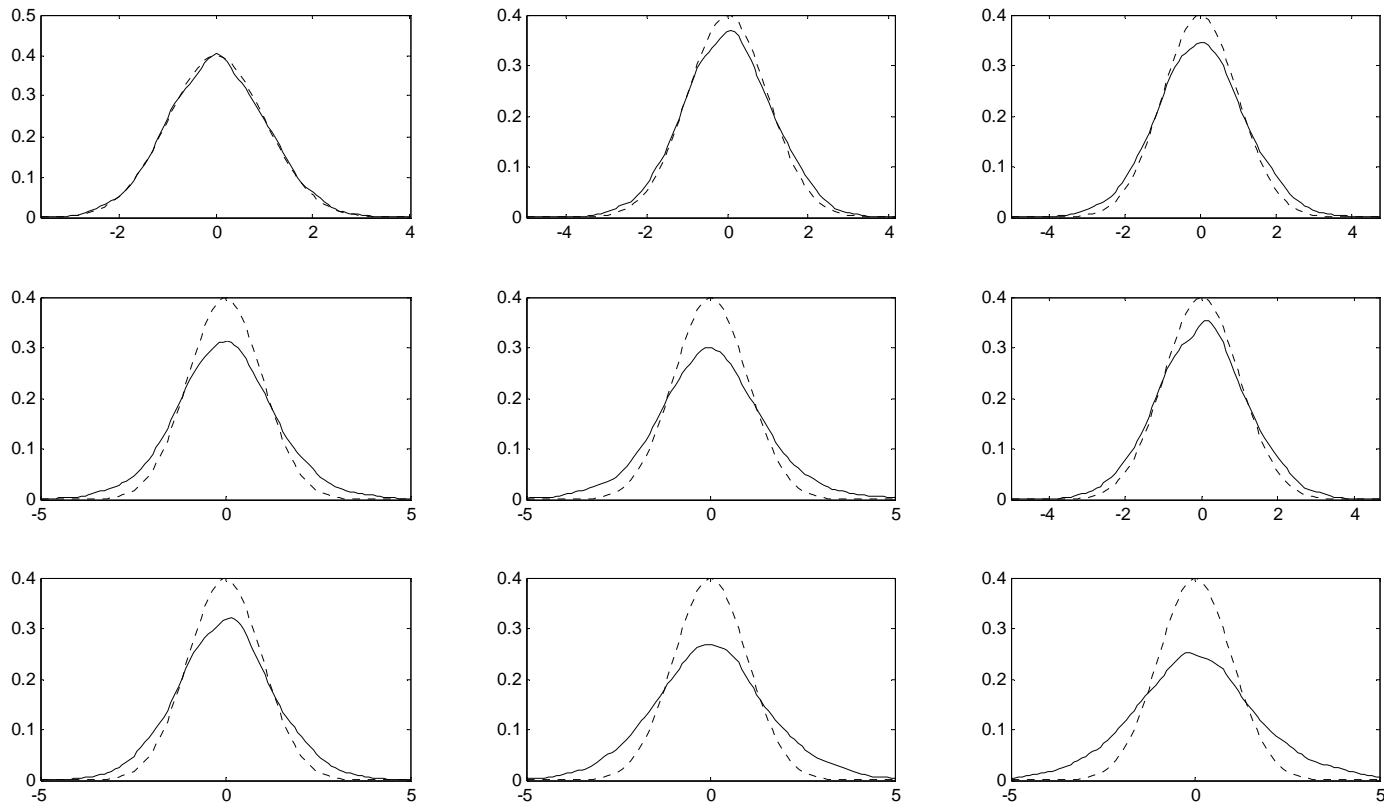


**Figure 10.**

Solid line: Estimated kernel density of studentised RV, using the estimated  $V$  as in Veraart (2007).

Dashed line: standard normal density.

Nine scenarios as described in Table 3.



**Figure 11.**  
 Solid line: Estimated kernel density of studentised RV, using the true V.  
 Dashed line: standard normal density.  
 Nine scenarios as described in Table 3.