

# Bayesian Nonparametric Estimation of Ex-post Variance\*

Jim Griffin<sup>†</sup>      Jia Liu<sup>‡</sup>      John M. Maheu<sup>§</sup>

This draft: September 2019

---

\*We are grateful for comments from Federico M. Bandi, an Associate Editor, two anonymous referees, Silvia Gonçalves and seminar participants at the 2016 NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics and the 2016 Rimini Conference in Economics and Finance. Maheu thanks the SSHRC of Canada for financial support.

<sup>†</sup>Department of Statistical Science, University College London, UK, j.griffin@ucl.ac.uk

<sup>‡</sup>Sobey School of Business, Saint Mary's University, Canada, Jia.Liu@smu.ca

<sup>§</sup>DeGroote School of Business, McMaster University, Canada, maheujm@mcmaster.ca

## Abstract

Variance estimation is central to many questions in finance and economics. Until now ex-post variance estimation has been based on infill asymptotic assumptions that exploit high-frequency data. This paper offers a new exact finite sample approach to estimating ex-post variance using Bayesian nonparametric methods. In contrast to the classical counterpart, the proposed method exploits pooling over high-frequency observations with similar variances. Bayesian nonparametric variance estimators under no noise, heteroskedastic and serially correlated microstructure noise are introduced and discussed. Monte Carlo simulation results show that the proposed approach can increase the accuracy of variance estimation. Applications to equity data and comparison with realized variance and realized kernel estimators are included.

Key words: Dirichlet process mixture, pooling, realized kernel, shrinkage

## 1 Introduction

Volatility is an indispensable quantity in finance and is a key input into asset pricing, risk management and portfolio management. In the last two decades, researchers have taken advantage of high-frequency data to estimate ex-post variance using intraperiod returns. Barndorff-Nielsen & Shephard (2002) and Andersen et al. (2003) formalized the idea of using high frequency data to measure the volatility of lower frequency returns. They show that realized variance (RV) is a consistent estimator of quadratic variation under ideal conditions. Unlike parametric models of volatility in which the model specification is important, RV is a *model free* estimate of quadratic variation in that it is valid under a wide range of spot volatility dynamics.<sup>1</sup>

RV provides an accurate measure of ex-post variance if there is no market microstructure noise. However, observed prices at high-frequency are inevitably contaminated by noise in reality and returns are no longer uncorrelated. In this case, RV is a biased and inconsistent estimator (Hansen & Lunde 2006, Aït-Sahalia et al. 2011). The impact of market microstructure noise on forecasting is explored in Aït-Sahalia & Mancini (2008) and Andersen et al. (2011).

Several different approaches have been proposed to estimating ex-post variance under microstructure noise. Zhou (1996) first introduced the idea of using a kernel-based method to estimate ex-post variance. Barndorff-Nielsen et al. (2008) formally discussed the realized kernel and showed how to use it in practice in a later paper (Barndorff-Nielsen et al. (2009)). Another approach is the subsampling method of Zhang et al. (2005). Hansen et al. (2008) showed how a time-series model can be used to filter out market microstructure to obtain corrected estimates of ex-post variance. A robust version of the predictive density of integrated volatility is derived in Corradi et al. (2009). Although bootstrap refinements are explored in Goncalves & Meddahi (2009) all distributional results from this literature rely on in-fill asymptotics.

Much of the literature has focused on the asymptotic properties of adaptations of realized variation which are robust to market microstructure noise. However, an argument can be made against the direct use of realized variation in the no noise situation if time between observations does not converge to zero. Realized variation is the sum of squared intraperiod returns and each component of that sum is an unbiased estimator of the corresponding intraperiod integrated volatility. It is well-understood that unbiased estimators based on one observations can be sub-optimal in terms of mean squared error and risk, *e.g.* Brown

& Zhao (2012). Shrinkage estimators which pool information from related estimates are one method to construct estimators with better properties. This suggests estimators of the integrated volatility with smaller mean squared error than realized variation can be constructed by summing shrunken estimates of the intraperiod integrated volatility.<sup>2</sup> When will there be substantial differences? Intuitively, shrinkage estimates work well when the unbiased estimators are noisy and information can be usefully pooled. If high frequency data had no noise, we would expect the difference to decrease as the sampling frequency increases (and the benefit of pooling will disappear asymptotically). In reality, high-frequency financial data includes noise and it is less clear that substantial differences will disappear asymptotically. In our simulation experiments, we find evidence that the difference in mean squared error persists at high frequency.

We use a Bayesian hierarchical approach to achieve shrinkage by pooling the information from related estimates. To our knowledge this idea has not been used in the estimation of volatility using high-frequency data. We assume that the intraperiod integrated volatility over short periods are exchangeable which implies that they can be modelled as conditionally independent and drawn from a prior distribution. The choice of distribution will have a strong effect on the form of pooling and so we choose to infer this distribution from the data using Bayesian nonparametric methods rather than choosing a parametric family (such as the generalized inverse Gaussian distribution). We model intraperiod returns according to a Dirichlet process mixture (DPM) model. This is a countably infinite mixture of distributions which facilitates the clustering of return observations into distinct groups sharing the same variance parameter.

Our proposed method benefits variance estimation in at least two aspects. First, the common values of intraperiod variance can be pooled into the same group leading to a more precise estimate. The pooling is done endogenously along with estimation of other model parameters. Second, the Bayesian nonparametric model delivers exact finite inference regarding ex-post variance or transformations such as the logarithm. As such, uncertainty around the estimate of ex-post volatility is readily available from the predictive density. Unlike the existing asymptotic theory which may give confidence intervals that contain negative values for variance, density intervals are always on the positive real line and can accommodate asymmetry. By extending key results in Hansen et al. (2008) we adapt the DPM mixture models to deal with returns contaminated with heteroskedastic noise and serially correlated noise.

Mykland & Zhang (2009) considered links between local parametric inference and high-frequency financial data analysis. Their approach assumes that quantities such as volatility are constant over blocks of returns and can lead to more efficient estimation and the definition of new estimators. Our method can be seen as a generalization of their blocked realized variance estimator using partitioning ideas from Bayesian nonparametrics by defining clusters rather than blocks of returns. Our approach endogenously finds these clusters and so does not restrict to clusters or blocks with a number of returns which are consecutive in time.

Monte Carlo simulation results show the Bayesian approach to be a very competitive alternative. Overall, pooling can lead to more precise estimates of ex-post variance and better coverage frequencies. These results are robust to different prior settings, irregularly spaced prices and tick time sampling. We show that the new variance estimators can be used with confidence and effectively recover both the average statistical features of daily ex-post variance as well as the time-series properties. Two applications to real world data with comparison to realized variance and kernel-based estimators are included.

This paper is organized as follows. The Bayesian nonparametric model, daily variance estimator and model estimation methods are discussed in Section 2. Section 3 extends the Bayesian nonparametric model to deal with heteroskedastic and serially correlated microstructure noise. The consistency of the estimator is considered in Section 4. Section 5 provides an extensive simulation and comparison of the estimators. Applications to IBM and S&P 500 ETF data are found in Section 6. Section 7 concludes followed by an appendix.

## 2 Bayesian Nonparametric Ex-post Variance Estimation

In this section, we introduce a Bayesian nonparametric ex-post volatility estimator. After defining the daily variance, conditional on the data, the discussion moves to the DPM model which provides the model framework of the proposed estimator. The approach discussed in this section deals with returns without microstructure noise and an estimator suitable for returns with microstructure noise is found in Section 3.

### 2.1 Model of High-frequency Returns

First, we consider the case with no market microstructure noise. We are interested in estimating the integrated volatility over fixed periods (which for simplicity will subsequently be called a day) using high-frequency intraday log returns. We will assume that there are  $n_t$  intraday log returns for the  $t$ -th day recorded at times  $\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,n_t}$  and which are denoted  $r_{t,1}, \dots, r_{t,n_t}$ . The model for each log-return is

$$r_{t,i} = \mu_t + \sigma_{t,i} z_{t,i}, \quad z_{t,i} \stackrel{iid}{\sim} N(0, 1), \quad i = 1, \dots, n_t, \quad (1)$$

where  $\mu_t$  is constant in day  $t$  and  $0 < \sigma_{t,i} < \infty$  for all  $i$ .<sup>3</sup> We make no assumptions on the stochastic process generating  $\sigma_{t,i}^2$ . For example, volatility may have jumps, undergo structural change or possess long memory. Given this assumed discrete data generating process we cannot distinguish between continuous and discrete (jumps) components as is commonly done in the literature (Barndorff-Nielsen & Shephard 2006). The daily return is

$$r_t = \sum_{i=1}^{n_t} r_{t,i} \quad (2)$$

and it follows, conditional on the unknown realized volatility path  $\mathcal{F}_t \equiv \{\sigma_{t,i}^2\}_{i=1}^{n_t}$ , the ex-post variance is

$$V_t \equiv \text{Var}(r_t | \mathcal{F}_t) = \sum_{i=1}^{n_t} \sigma_{t,i}^2. \quad (3)$$

In our Bayesian setting  $V_t$  is the target to estimate conditional on the data  $\{r_{t,i}\}_{i=1}^{n_t}$ .

### 2.2 A Bayesian Model with Pooling

In this section we discuss a nonparametric prior for the model of (1) that allows for pooling over common values of  $\sigma_{t,i}^2$ . The Dirichlet process mixture model (DPM) is a Bayesian

nonparametric mixture model that has been used in density estimation and for modeling unknown hierarchical effects among many other applications. A key advantage of the model is that it naturally incorporates parameter pooling.

Our nonparametric model has the following hierarchical form

$$r_{t,i} | \mu_t, \sigma_{t,i} \stackrel{iid}{\sim} N(\mu_t, \sigma_{t,i}^2), \quad i = 1, \dots, n_t, \quad (4)$$

$$\sigma_{t,i}^2 | G_t \stackrel{iid}{\sim} G_t, \quad (5)$$

$$G_t | G_{0,t}, \alpha_t \sim \text{DP}(\alpha_t, G_{0,t}), \quad (6)$$

$$G_{0,t} \equiv \text{IG}(v_{0,t}, s_{0,t}), \quad (7)$$

where the base measure is the inverse-gamma distribution denoted as  $\text{IG}(v, s)$ , which has a mean of  $s/(v - 1)$  for  $v > 1$ . The return mean  $\mu_t$  is assumed to be a constant over  $i$ .

The Dirichlet process was formally introduced by Ferguson (1973) and is a distribution over distributions. Note that each day assumes an independent DPM model and is indexed by  $t$ . We do not pursue pooling over  $G_t$ . A draw from a  $\text{DP}(\alpha_t, G_{0,t})$  is an almost surely discrete distribution which is centered around the base distribution  $G_{0,t}$ , in the sense that  $E[G_t(B)] = G_{0,t}(B)$  for any set  $B$ . The concentration parameter  $\alpha_t > 0$  governs how closely a draw  $G_t$  resembles  $G_{0,t}$ . Larger values of  $\alpha_t$  lead to  $G_t$  being closer to  $G_{0,t}$ . Since the realisation  $G_t$  are discrete, a sample from  $\sigma_{t,i}^2 | G_t \sim G_t$  has a positive probability of repeated values. This has led the use of DPM's for clustering problems. If  $K_n$  is the number of distinct values in a sample of size  $n$ , then  $E[K_n] \approx \alpha_t \log(1 + n/\alpha_t)$ . Therefore, the number of distinct values grows logarithmically with sample size and larger values of  $\alpha_t$  will tend to lead to more distinct values. In fact, as  $\alpha_t \rightarrow \infty$ ,  $G_t \rightarrow G_{0,t}$  which implies that every  $r_{t,i}$  has a unique  $\sigma_{t,i}^2$  drawn from centering distribution. In this model, the inverse gamma distribution is used as the centering base measure as this is the standard conjugate choice and leads to relatively simple computational schemes for inference.

In the case of  $\alpha_t \rightarrow \infty$ , there is no pooling and we have a setting very close to the classical counterpart discussed above. However, for finite  $\alpha_t$ , pooling can take place. The other extreme is complete pooling for  $\alpha_t \rightarrow 0$  in which there is one common variance shared by all observations such that  $\sigma_{t,i}^2 = \sigma_{t,1}^2, \forall i$ . Since  $\alpha_t$  plays an important role in pooling we place a prior on it and estimate it along with the other model parameters for each day.

A stick breaking representation (Sethuraman 1994) of the DPM in (5) is given as follows.<sup>4</sup>

$$p(r_{t,i} | \mu_t, \Psi_t, w_t) = \sum_{j=1}^{\infty} w_{t,j} N(r_{t,i} | \mu_t, \psi_{t,j}^2), \quad (8)$$

$$w_{t,j} = v_{t,j} \prod_{l=1}^{j-1} (1 - w_{t,l}), \quad (9)$$

$$v_{t,j} \stackrel{iid}{\sim} \text{Beta}(1, \alpha_t), \quad (10)$$

where  $N(\cdot | \cdot, \cdot)$  denotes the density of the normal distribution,  $\Psi_t = \{\psi_{t,1}^2, \psi_{t,2}^2, \dots\}$  is the set of unique values of  $\sigma_{t,i}^2$ ,  $w_t = \{w_{t,1}, w_{t,2}, \dots\}$  and  $w_{t,j}$  is the weight associated with the  $j^{\text{th}}$  component. This formulation of the model facilitates posterior sampling which is discussed in the next section.

Since our focus is on intraday returns and the number of observations in a day can be small, especially for lower frequencies such as 5-minute. Therefore, the prior should be chosen

carefully. It is straightforward to show that the prior predictive distribution of  $\sigma_{t,i}^2$  is  $G_{0,t}$ . For  $\sigma_{t,i}^2 \sim \text{IG}(v_{0,t}, s_{0,t})$ , the mean and variance of  $\sigma_{t,i}^2$  are

$$\mathbb{E}(\sigma_{t,i}^2) = \frac{s_{0,t}}{v_{0,t} - 1} \quad \text{and} \quad \text{var}(\sigma_{t,i}^2) = \frac{s_{0,t}^2}{(v_{0,t} - 1)^2(v_{0,t} - 2)}. \quad (11)$$

Solving the two equations, the values of  $v_{0,t}$  and  $s_{0,t}$  are given by

$$v_{0,t} = \frac{[\mathbb{E}(\sigma_{t,i}^2)]^2}{\text{var}(\sigma_{t,i}^2)} + 2 \quad \text{and} \quad s_{0,t} = \mathbb{E}(\sigma_{t,i}^2)(v_{0,t} - 1). \quad (12)$$

We use sample statistics  $\widehat{\text{var}}(r_{t,i})$  and  $\widehat{\text{var}}(r_{t,i}^2)$  calculated with three days intraday returns (day  $t - 1$ , day  $t$ , and day  $t + 1$ ) to set the values of  $\mathbb{E}(\sigma_{t,i}^2)$  and  $\text{var}(\sigma_{t,i}^2)$ , then use equation (12) to find  $v_{0,t}$  and  $s_{0,t}$ . A shrinkage prior  $\text{N}(0, v^2)$  is used for  $\mu_t$  since  $\mu_t$  is expected to be close to zero. The prior variance of  $\mu_t$  is adjusted according to data frequency:  $v^2 = \zeta^2/n_t$  where  $n_t$  is the number of intraday returns. Finally,  $\alpha_t \sim \text{Gamma}(a, b)$ .

For a finite dataset  $i = 1, \dots, n_t$  our target is the following posterior moment

$$E[V_t | \{r_{t,i}\}_{i=1}^{n_t}] = E \left[ \sum_{i=1}^{n_t} \sigma_{t,i}^2 \middle| \{r_{t,i}\}_{i=1}^{n_t} \right]. \quad (13)$$

Note that the posterior mean of  $V_t$  can also be considered as the posterior mean of realized variance,  $RV_t = \sum_{i=1}^{n_t} r_{t,i}^2$  assuming  $\mu_t$  is small. As such,  $RV_t$  treats each  $\sigma_{t,i}^2$  as separate and corresponds to no pooling.

Mykland & Zhang (2009) discuss the use of blocks of high frequency data in volatility estimation. Our method can be seen as a generalization of Mykland & Zhang (2009). We allow returns with the same variance to form groups flexibly and do not impose the restriction that the returns in one group are consecutive in time. Another distinction is that our approach allows the group size to vary over clusters and be determined endogenously, while Mykland & Zhang (2009) has one fixed block size for all clusters preset by the econometrician. Furthermore, unlike standard blocking, the proposed method is invariance to return permutations since the DPM model assumes exchangeable data.

## 2.3 Model Estimation

Estimation relies on Markov chain Monte Carlo (MCMC) techniques. We apply the slice sampler of Kalli et al. (2011), along with Gibbs sampling to estimate the DPM model. The slice sampler provides an elegant way to deal with the infinite states in (8). It introduces an auxiliary variable  $u_{t,1:n_t} = \{u_{t,1}, \dots, u_{t,n_t}\}$  that randomly truncates the state space to a finite set at each MCMC iteration but marginally delivers draws from the desired posterior.

The joint distribution of  $r_{t,i}$  and the auxiliary variable  $u_{t,i}$  is given by

$$f(r_{t,i}, u_{t,i} | w_t, \mu_t, \Psi_t) = \sum_{j=1}^{\infty} \mathbb{1}(u_{t,i} < w_{t,j}) \text{N}(r_{t,i} | \mu_t, \psi_{t,j}^2), \quad (14)$$

and integrating out  $u_{t,i}$  recovers (8).

It is convenient to rewrite the model in terms of a latent state variable  $s_{t,i} \in \{1, 2, \dots\}$  that maps each observation to an associated component and parameter  $\sigma_{t,i}^2 = \psi_{t,s_{t,i}}^2$ . Observations

with a common state share the same variance parameter. For finite dataset the number of states (clusters) is finite and ordered from  $1, \dots, K$ . Note that the number of clusters  $K$ , is not a fixed value over the MCMC iterations. A new cluster with variance  $\psi_{t,K+1}^2 \sim G_{0,t}$  can be created if existing clusters do not fit that observation well and clusters sharing a similar variance can be merged into one.

The joint posterior is

$$p(\mu_t) \prod_{j=1}^K [p(\psi_{t,j}^2)] p(\alpha_t) \prod_{i=1}^{n_t} \mathbb{1}(u_{t,i} < w_{t,s_{t,i}}) \text{N}(r_{t,i} | \mu_t, \psi_{t,s_{t,i}}^2). \quad (15)$$

Each MCMC iteration contains the following sampling steps.

1.  $\pi(\mu_t | r_{t,1:n_t}, \{\psi_{t,j}^2\}_{j=1}^K, s_{t,1:n_t}) \propto p(\mu_t) \prod_{i=1}^{n_t} p(r_{t,i} | \mu_t, \psi_{t,s_{t,i}}^2)$ .
2.  $\pi(\psi_{t,j}^2 | r_{t,1:n_t}, s_{t,1:n_t}, \mu_t) \propto p(\psi_{t,j}^2) \prod_{t:s_{t,i}=j} p(r_{t,i} | \mu_t, \psi_{t,j}^2)$  for  $j = 1, \dots, K$ .
3.  $\pi(v_{t,j} | s_{t,1:n_t}) \propto \text{Beta}(v_{t,j} | a_{t,j}, b_{t,j})$  with  $a_{t,j} = 1 + \sum_{i=1}^{n_t} \mathbb{1}(s_{t,i} = j)$  and  $b_{t,j} = \alpha_t + \sum_{i=1}^{n_t} \mathbb{1}(s_{t,i} > j)$  and update  $w_{t,j} = v_{t,j} \prod_{l < j} (1 - v_{t,l})$  for  $j = 1, \dots, K$ .
4.  $\pi(u_{t,i} | w_{t,i}, s_{t,1:n_t}) \propto \mathbb{1}(0 < u_{t,i} < w_{t,s_{t,i}})$ .
5. Find the smallest  $K$  such that  $\sum_{j=1}^K w_{t,j} > 1 - \min(u_{t,1:n_t})$ .
6.  $\pi(s_{t,i} | r_{t,1:n_t}, s_{t,1:n_t}, \mu_t, \{\psi_{t,j}^2\}_{j=1}^K, u_{t,1:n_t}, K) \propto \sum_{j=1}^K \mathbb{1}(u_{t,i} < w_{t,j}) p(r_{t,i} | \mu_t, \psi_{t,j}^2)$  for  $i = 1, \dots, n_t$ .
7.  $\pi(\alpha_t | K) \propto p(\alpha_t) p(K | \alpha_t)$ .

In the first step  $\mu_t$  is common to all returns and this is a standard Gibbs step given the conjugate prior. Step 2 is a standard Gibbs step for each variance parameter  $\psi_{t,j}^2$  based on the data assigned to cluster  $j$ . The remaining steps are standard for slice sampling of DPM models. In 7,  $\alpha_t$  is sampled based on Escobar & West (1994).

Steps 1-7 give one iteration of the posterior sampler. After dropping a suitable burn-in amount,  $M$  additional samples are collected,  $\{\theta^{(m)}\}_{m=1}^M$ , where  $\theta = \{\mu_t, \psi_{t,1}^2, \dots, \psi_{t,K}^2, s_{t,1:n_t}, \alpha_t\}$ . Posterior moments of interest can be estimated from sample averages of the MCMC output.

## 2.4 Ex-post Variance Estimator

Conditional on the parameter vector  $\theta$  the estimate of  $V_t$  is

$$E[V_t | \theta] = \sum_{i=1}^{n_t} \sigma_{t,s_i}^2. \quad (16)$$

The posterior mean of  $V_t$  is obtained by integrating out all parameter and distributional uncertainty.  $E[V_t | \{r_{t,i}\}_{i=1}^{n_t}]$  is estimated as

$$\hat{V}_t = \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^{n_t} \sigma_{t,i}^{2(m)}, \quad (17)$$

where  $\sigma_{t,i}^{2(m)} = \psi_{t,s_{t,i}}^{2(m)}$ . Similarly other features of the posterior distribution of  $V_t$  can be obtained. For instance, a  $(1-\alpha)$  probability density interval for  $V_t$  is the quantiles of  $\sum_{i=1}^{n_t} \sigma_{t,s_{t,i}}^2$  associated with probabilities  $\alpha/2$  and  $(1 - \alpha/2)$ . Conditional on the model and prior these are exact finite sample estimates, in contrast to the classical estimator which relies on infill asymptotics to derived confidence intervals.

If  $\log(V_t)$  is the quantity of interest, the estimator of  $E[\log(V_t)|\{r_{t,i}\}_{i=1}^{n_t}]$  is given as

$$\widehat{\log(V_t)} = \frac{1}{M} \sum_{m=1}^M \log \left( \sum_{i=1}^{n_t} \sigma_{t,i}^{2(m)} \right). \quad (18)$$

As before, quantile estimates of the posterior of  $\log(V_t)$  can be estimated from the MCMC output.

### 3 Bayesian Estimator Under Microstructure Error

An early approach to deal with market microstructure noise was to prefilter with a time-series model (Andersen, Bollerslev, Diebold & Ebens 2001, Bollen & Inder 2002, Maheu & McCurdy 2002). Hansen et al. (2008) shows that prefiltering results in a bias to realized variance that can be easily corrected. We employ these insights into moving average specifications to account for noisy high-frequency returns. A significant difference is that we allow for heteroskedasticity in the noise process.

#### 3.1 DPM-MA(1) Model

The existence of microstructure noise turns the intraday return process into an autocorrelated process. First consider the case in which the error is white noise:

$$\tilde{p}_{t,i} = p_{t,i} + \epsilon_{t,i}, \quad \epsilon_{t,i} \sim N(0, \omega_{t,i}^2), \quad (19)$$

where  $\tilde{p}_{t,i}$  denotes the observed log-price with error,  $p_{t,i}$  is the unobserved fundamental log-price and  $\omega_{t,i}^2$  is the heteroskedastic noise variance.

Given this structure it can be shown that the returns series  $\tilde{r}_{t,i} = \tilde{p}_{t+1,i} - \tilde{p}_{t,i}$  has non-zero first order autocorrelation but zero higher order autocorrelation. That is  $\text{cov}(\tilde{r}_{t,i+1}, \tilde{r}_{t,i}) = -\omega_{t,i}^2$  and  $\text{cov}(\tilde{r}_{t,i+j}, \tilde{r}_{t,i}) = 0$  for  $j \geq 2$ . This suggests a moving average model of order one.<sup>5</sup>

Combining MA(1) parameterization with our Bayesian nonparametric framework yields the DPM-MA(1) models.

$$\tilde{r}_{t,i} | \mu_t, \theta_t, \delta_{t,i}^2 = \mu_t + \theta_t \eta_{t,i-1} + \eta_{t,i}, \quad \eta_{t,i} \sim N(0, \delta_{t,i}^2) \quad (20)$$

$$\delta_{t,i}^2 | G_t \sim G_t, \quad (21)$$

$$G_t | G_{0,t}, \alpha_t \sim \text{DP}(\alpha_t, G_{0,t}), \quad (22)$$

$$G_{0,t} \equiv \text{IG}(v_{0,t}, s_{0,t}). \quad (23)$$

The noise terms are heteroskedastic. Note that the mean of  $r_{t,i}$  is not a constant term but a moving average term. The MA parameter  $\theta_t$  is constant for  $i$  but will change with the day  $t$ . The prior is  $\theta_t \sim N(m_\theta, v_\theta^2) \mathbb{1}_{\{|\theta_t| < 1\}}$  in order to make the MA model invertible. The error term  $\eta_{t,0}$  is assumed to be zero. Other model settings remain the same as the DPM illustrated in Section 2. Later we show how estimates from this specification can be used to recover an estimate of the ex-post variance  $V_t$  of the true return process.



## 3.2 DPM-MA(q) Model

For lower sampling frequencies, such as 1 minute or more, first order autocorrelation is the main effect from market microstructure. As such, the MA(1) model will be sufficient for many applications. However, at higher sampling frequencies, the dependence may be stronger. To allow for a more complex effect on returns from the noise process consider the MA(q-1) noise affecting returns,

$$\tilde{p}_{t,i} = p_{t,i} + \epsilon_{t,i} - \rho_1 \epsilon_{t,i-1} - \dots - \rho_{q-1} \epsilon_{t,i-q+1}, \quad \epsilon_{t,i} \sim N(0, \omega_{t,i}^2). \quad (24)$$

For returns, this leads to the following DPM-MA(q) model,

$$\tilde{r}_{t,i} | \mu_t, \{\theta_{t,j}\}_{j=1}^q, \delta_{t,i}^2 = \mu_t + \sum_{j=1}^q \theta_{t,j} \eta_{t,i-j} + \eta_{t,i}, \quad \eta_{t,i} \sim N(0, \delta_{t,i}^2) \quad (25)$$

$$\delta_{t,i}^2 | G_t \sim G_t, \quad (26)$$

$$G_t | G_{0,t}, \alpha_t \sim \text{DP}(\alpha_t, G_{0,t}), \quad (27)$$

$$G_{0,t} \equiv \text{IG}(v_{0,t}, s_{0,t}). \quad (28)$$

The joint prior of  $(\theta_{t,1}, \dots, \theta_{t,q})$  is  $N(M_\Theta, V_\Theta) \mathbb{1}_{\{\Theta\}}^6$  and  $(\eta_{t,0}, \dots, \eta_{t,-(q-1)}) = (0, \dots, 0)$ .

## 3.3 Model Estimation

We discuss the estimation of DPM-MA(1) model and the approach can be easily extended to the DPM-MA(q). The main difference in this model is that the conditional mean parameters  $\mu_t$  and  $\theta_t$  require a Metropolis-Hasting (MH) step to sample their conditional posteriors. The remaining MCMC steps are essentially the same. As before, let  $\psi_{t,i}^2$  denote the unique values of  $\delta_{t,j}^2$  then each MCMC iteration samples from the following conditional distributions.

1.  $\pi(\mu_t | \tilde{r}_{t,1:n_t}, \{\psi_{t,j}^2\}_{j=1}^K, \theta_t, s_{t,1:n_t}) \propto p(\mu_t) \prod_{i=1}^{n_t} N(\tilde{r}_{t,i} | \mu_t + \theta_t \eta_{t,i-1}, \psi_{t,s_{t,i}}^2)$ .
2.  $\pi(\theta_t | \tilde{r}_{t,1:n_t}, \mu_t, \{\psi_{t,j}^2\}_{j=1}^K, s_{t,1:n_t}^t) \propto p(\theta_t) \prod_{i=1}^{n_t} p(\tilde{r}_{t,i} | \mu_t + \theta_t \eta_{t,i-1}, \psi_{t,s_{t,i}}^2)$ .
3.  $\pi(\psi_{t,j}^2 | \tilde{r}_{t,1:n_t}, \mu_t, \theta_t, s_{t,1:n_t}) \propto p(\psi_{t,j}^2) \prod_{t:s_{t,i}=j} p(\tilde{r}_{t,i} | \mu_t + \theta_t \epsilon_{t,i-1}, \psi_{t,j}^2)$  for  $j = 1, \dots, K$ .
4.  $\pi(v_{t,j} | s_{t,1:n_t}) \propto \text{Beta}(v_{t,j} | a_{t,j}, b_{t,j})$  with  $a_{t,j} = 1 + \sum_{i=1}^{n_t} \mathbb{1}(s_{t,i} = j)$  and  $b_{t,j} = \alpha_t + \sum_{i=1}^{n_t} \mathbb{1}(s_{t,i} > j)$  and update  $w_{t,j} = v_{t,j} \prod_{l < j} (1 - v_{t,l})$  for  $j = 1, \dots, K$ .
5.  $\pi(u_{t,i} | w_{t,i}, s_{t,1:n_t}) \propto \mathbb{1}(0 < u_{t,i} < w_{t,s_{t,i}})$  for  $i = 1, \dots, n_t$ .
6. Find the smallest  $K$  such that  $\sum_{j=1}^K w_{t,j} > 1 - \min(u_{t,1:n_t})$ .
7.  $\pi(s_{t,i} | \tilde{r}_{t,1:n_t}, s_{t,1:n_t}, \mu_t, \theta_t, \{\psi_{t,j}^2\}_{j=1}^K, u_{t,1:n_t}, K) \propto \sum_{j=1}^K \mathbb{1}(u_{t,i} < w_{t,j}) N(\tilde{r}_{t,i} | \mu_t + \theta_t \eta_{t,i-1}, \psi_{t,j}^2)$  for  $i = 1, \dots, n_t$ .
8.  $\pi(\alpha_t | K) \propto p(\alpha_t) p(K | \alpha_t)$ .

In steps 1 and 2 the likelihood requires the sequential calculation of the lagged error as  $\eta_{t,i-1} = \tilde{r}_{t,i-1} - \mu_t - \theta_t \eta_{t,i-2}$  which precludes a Gibbs sampling step. Therefore,  $\mu_t$  and  $\theta_t$  are sampled using a MH with a random walk proposal. The proposal is calibrated to achieve an acceptance rate between 0.3 and 0.5.

### 3.4 Ex-post Variance Estimator under Microstructure Error

Hansen et al. (2008) showed that prefiltering with an MA model results in a bias in the RV estimator.<sup>7</sup> In the Appendix it is shown that the Hansen et al. (2008) bias correction provides an accurate adjustment to our Bayesian estimator in the context of heteroskedastic noise. From the DPM-MA(1) model the posterior mean of  $V_t$  under independent microstructure error is

$$\hat{V}_{t,MA(1)} = \frac{1}{M} \sum_{m=1}^M (1 + \theta_t^{(m)})^2 \sum_{i=1}^{n_t} \delta_{t,i}^{2(m)}, \quad (29)$$

where  $\delta_{t,i}^{2(m)} = \psi_{t,s_{t,i}}^{2(m)}$ . The log of  $V_t$ , square-root of  $V_t$  and density intervals can be estimated as the Bayesian nonparametric ex-post variance estimator without microstructure error.

In the case of higher autocorrelation the DPM-MA(q) model adjusted posterior estimate of  $V_t$  is

$$\hat{V}_{t,MA(q)} = \frac{1}{M} \sum_{m=1}^M \left( 1 + \sum_{j=1}^q \theta_{t,j}^{(m)} \right)^2 \sum_{i=1}^{n_t} \delta_{t,i}^{2(m)}. \quad (30)$$

Next we consider simulation evidence on these estimators.

## 4 Consistency

Each of the previous discussed estimators (posterior means) for integrated volatility (3) can be fairly easily shown to be consistent estimators as the sampling frequency increases. The posterior mean of  $V_t$  can be shown to be equal to a consistent estimator plus a bias term that goes to zero in probability as  $n_t \rightarrow \infty$ . We provide the proof for the case with no market microstructure noise.

**Theorem 1.** *Suppose that  $p$  is an arbitrage-free price process with zero mean, that  $\sup(\tau_{t,j+1} - \tau_{t,j}) \rightarrow 0$  for  $n_t \rightarrow \infty$  and that  $s_{0,n_t} = O(n_t^{-\alpha})$  for  $\alpha > 0$  then  $E[V_t | \{r_{t,i}\}_{i=1}^{n_t}]$  is a consistent estimator of the integrated volatility.*

See the Appendix for the proof.

A similar argument can be used for the MA processes with the residuals from the MA process replacing the returns. (Hansen et al. 2008) argue that scaling  $RV_t$  avoids the inconsistency of realized volatility under market microstructure noise.

## 5 Simulation Results

### 5.1 Data Generating Process

We consider four commonly used data generating processes (DGPs) in the literature. The first one is the GARCH(1,1) diffusion, introduced by Andersen & Bollerslev (1998). The log-price follows

$$dp(t) = \mu dt + \sigma(t) dW_p(t), \quad (31)$$

$$d\sigma^2(t) = \alpha(\beta - \sigma^2(t))dt + \gamma\sigma^2(t)dW_\sigma(t). \quad (32)$$

where  $W_p(t)$  and  $W_\sigma(t)$  are two independent Wiener processes. The values of parameters follow Andersen & Bollerslev (1998) and are  $\mu = 0.03$ ,  $\alpha = 0.035$ ,  $\beta = 0.636$  and  $\gamma = 0.144$ , which were estimated using foreign exchange data.

Following Huang & Tauchen (2005), the second and third DGP are a one factor stochastic volatility diffusion (SV1F) and one factor stochastic volatility diffusion with jumps (SV1FJ). SV1F is given by

$$dp(t) = \mu dt + \exp(\beta_0 + \beta_1 v(t)) dW_p(t), \quad (33)$$

$$dv(t) = \alpha v(t) dt + dW_v(t) \quad (34)$$

and the price process for SV1FJ is

$$dp(t) = \mu dt + \exp(\beta_0 + \beta_1 v(t)) dW_p(t) + dJ(t), \quad (35)$$

where  $\text{corr}(dW_p(t), dW_v(t)) = \rho$ , and  $J(t)$  is a Poisson process with jump intensity  $\lambda$  and jump size  $\delta \sim N(0, \sigma_J^2)$ . We adopt the parameter settings from Huang & Tauchen (2005) and set  $\mu = 0.03$ ,  $\beta_0 = 0.0$ ,  $\beta_1 = 0.125$ ,  $\alpha = -0.1$ ,  $\rho = -0.62$ ,  $\lambda = 0.014$  and  $\sigma_J^2 = 0.5$ .

The final DGP is the two factor stochastic volatility diffusion (SV2F) from Chernov et al. (2003) and Huang & Tauchen (2005).<sup>8</sup>

$$dp(t) = \mu dt + s \exp(\beta_0 + \beta_1 v_1(t) + \beta_2 v_2(t)) dW_p(t), \quad (36)$$

$$dv_1(t) = \alpha_1 v_1(t) dt + dW_{v_1}(t), \quad (37)$$

$$dv_2(t) = \alpha_2 v_2(t) dt + (1 + \psi v_2(t)) dW_{v_2}(t), \quad (38)$$

where  $\text{corr}(dW_p(t), dW_{v_1}(t)) = \rho_1$  and  $\text{corr}(dW_p(t), dW_{v_2}(t)) = \rho_2$ . The parameter values in SV2F are  $\mu = 0.03$ ,  $\beta_0 = -1.2$ ,  $\beta_1 = 0.04$ ,  $\beta_2 = 1.5$ ,  $\alpha_1 = -0.00137$ ,  $\alpha_2 = -1.386$ ,  $\psi = 0.25$  and  $\rho_1 = \rho_2 = -0.3$ , which are from Huang & Tauchen (2005).

Data is simulated using a basic Euler discretization at 1-second frequency for the four DGPs. Assuming the length of daily trading time is 6.5 hours (23400 seconds), we first simulate the log price level every second. After this we compute the 5-minute, 1-minute, 30-second and 10-second intraday returns by taking the difference every 300, 60, 30, 10 steps, respectively. The initial volatility level, such as  $v_{1t}$  and  $v_{2t}$  in SV2F, at day  $t$  is set equal to the last volatility value at previous day,  $t-1$ .  $T = 5000$  days of intraday returns are simulated using the four DGPs and used to report sampling properties of the volatility estimators. In each case, to remove dependence on the startup conditions 500 initial days are dropped from the simulation.

### 5.1.1 Independent Noise

Following Barndorff-Nielsen et al. (2008), log-prices with independent noise are simulated as follows

$$\begin{aligned} \tilde{p}_{t,i} &= p_{t,i} + \epsilon_{t,i}, \\ \epsilon_{t,i} &\sim N(0, \sigma_\omega^2), \\ \sigma_\omega^2 &= \xi^2 \text{var}(r_t). \end{aligned} \quad (39)$$

The error term is added to the log-prices simulated from the 4 DGPs every second. The variance of microstructure error is proportional to the daily variance calculated using the pure daily returns. We set the noise-to-signal ratio  $\xi^2 = 0.001$ , which is the same value used in Barndorff-Nielsen et al. (2008) and close to the value in Bandi & Russell (2008).

### 5.1.2 Dependent Noise

Following Hansen et al. (2008), we consider the simulation of log-prices with dependent noise as follows,

$$\begin{aligned}
 \tilde{p}_{t,i} &= p_{t,i} + \epsilon_{t,i}, \\
 \epsilon_{t,i} &\sim N(\mu_{\epsilon_{t,i}}, \sigma_\omega^2), \\
 \mu_{\epsilon_{t,i}} &= \sum_{l=1}^{\phi} \left(1 - \frac{l}{\phi}\right) (p_{t,i-l} - p_{t,i-1-l}), \\
 \sigma_\omega^2 &= \xi^2 \text{var}(r_t),
 \end{aligned} \tag{40}$$

where  $\phi = 20$ , which makes the error term correlated with returns in the past 20 seconds (steps). If past returns are positive (negative) the noise term tends to be positive (negative). All other settings, such as  $\sigma_\omega^2$  and  $\xi^2$ , are the same as in the independent error case.

## 5.2 True Volatility and Comparison Criteria

The realized variance ( $RV$ ) by Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2002), the flat-top realized kernel ( $RK^F$ ) by Barndorff-Nielsen et al. (2008) and the non-negative realized kernel ( $RK^N$ ) by Barndorff-Nielsen et al. (2011) serve as the benchmarks for comparison. Section 8.1 in the appendix provides a brief review of those estimators.

We assess the ability of several ex-post variance estimators to estimate the daily quadratic variation ( $QV_t$ ) from the four data generating processes.  $QV_t$  is estimated as the summation of the squared intraday pure returns at the highest frequency (1 second)

$$\sigma_t^2 \equiv \sum_{i=1}^{23400} r_{t,i}^2. \tag{41}$$

The competing ex-post daily variance estimators, generically labelled  $\hat{\sigma}_t^2$ , are compared based on the root mean squared errors (RMSE), and bias defined as

$$\text{RMSE}(\hat{\sigma}_t^2) = \sqrt{\frac{1}{T} \sum_{t=1}^T (\hat{\sigma}_t^2 - \sigma_t^2)^2}, \tag{42}$$

$$\text{Bias}(\hat{\sigma}_t^2) = \frac{1}{T} \sum_{t=1}^T (\hat{\sigma}_t^2 - \sigma_t^2). \tag{43}$$

The coverage probability estimates report the frequency that the confidence intervals or density intervals from the Bayesian nonparametric estimators contain the true ex-post variance,  $\sigma_t^2$ . The 95% confidence intervals of  $RV_t$ ,  $RK_t^F$  and  $RK_t^N$  rely on the asymptotic distribution, which are provided in equation (49), (52) and (56). We take the bias into account to compute the 95% confidence interval using  $RK_t^N$ .

The estimation of integrated quarticity is crucial in determining the confidence interval for the realized kernels. We consider two versions of quarticity, one is to use the true (infeasible)  $IQ_t$  which is calculated as

$$IQ_t^{\text{true}} = 23400 \sum_{i=1}^{23400} \sigma_{t,i}^4, \tag{44}$$

where  $\sigma_{t,i}^2$  refers to spot variance simulated at the highest frequency. The other method is to estimate  $IQ_t$  using the tri-power quarticity estimator, see formula (54). The confidence interval based on  $IQ_t^{\text{true}}$  is the infeasible case and the confidence interval calculated using  $TPQ_t$  is the feasible case.

For each day 5000 MCMC draws are collected after 1000 burn-in to compute the Bayesian posterior quantities. A 0.95 density interval is the 0.025 and 0.975 sample quantiles of MCMC draws of  $\sum_{i=1}^{n_t} \sigma_{t,i}^2$ , respectively.

### 5.3 No Microstructure Noise

In Table 2,  $\hat{V}_t$  has slightly smaller RMSE in 12 out of the 16 categories. A paired t-test shows most of these differences in MSE are significant as well. For example, for the 5-minute data  $\hat{V}_t$  reduces the RMSE by over 5% for the SV2F data. This is remarkable given that  $RV_t$  is the gold standard in the no noise setting. Bias and coverage probabilities (not displayed) for 95% confidence intervals of  $RV_t$  and 0.95 density intervals of  $\hat{V}_t$  show both estimators to perform well. Under no microstructure noise, the Bayesian nonparametric estimator is competitive with the classical counterpart  $RV_t$ .  $\hat{V}_t$  offers smaller estimation error and better finite sample results than  $RV_t$  when the data frequency is low. Performance of  $RV_t$  and  $\hat{V}_t$  both improve as the sampling frequency increases.

The comparison between the Bayes nonparametric estimator and the blocked  $RV^9$  of Mykland & Zhang (2009) is also considered. Table 2 reports RMSE of blocked  $RV$  with block size being  $n_t^{3/4}$ . The RMSE of  $\hat{V}_t$  remains the lowest in 12 out of the 16 cases.

A robustness analysis is conducted to check how sensitive the results are to the selection of priors. Different sets of hyperparameters of  $\mu_t$  and  $\alpha$  are considered and calibration of the prior of  $\sigma_{t,i}^2$  based on only one day of data (Equation (12)). Table 3 summarizes RMSE of  $\hat{V}_t$  under alternative priors for SV2F data (see entries of  $v_{0,t}^{1d}$ ,  $s_{0,t}^{1d}$  for one day of prior calibration). None of the result changes more than 1% under new priors and  $\hat{V}_t$  consistently outperform  $RV_t$  in 5-minute, 1-minute and 30-second categories.

We also include an analysis to check if the benefits of pooling persists given irregularly spaced returns. Following Barndorff-Nielsen et al. (2011), arrival times of observed prices are simulated from a Poisson process. Table 4 shows the Bayes nonparametric estimator  $\hat{V}_t$  has a lower RMSE compared with  $RV_t$  for this irregular-spaced data generating process.

### 5.4 Independent and Dependent Microstructure Noise

In this section we compare  $RV_t$ ,  $RK_t^F$ ,  $\hat{V}_t$  and  $\hat{V}_{t,MA(1)}$  with independent microstructure noise. Table 5 shows the RMSE of the various estimators for different sampling frequencies and DGPs.  $RV_t$  and  $\hat{V}_t$  produce smaller errors in estimating  $\sigma_t^2$  than  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  for 5-minute data. However, increasing the sampling frequency results in a larger bias from the microstructure noise. As such,  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  are more accurate as the data frequency increases. Compared to  $RK_t^F$ ,  $\hat{V}_{t,MA(1)}$  has a smaller RMSE in all cases, except for 30-second and 10-second SV2F return.

As can be seen in Table 6,  $\hat{V}_{t,MA(1)}$  has the best finite sample coverage among all the alternatives except for the SV2F data. For example, the coverage probabilities of 0.95 density intervals are always within 0.5% from the truth. Note that the density intervals are trivial to obtain from the MCMC output and do not require the calculation  $IQ_t$ . The coverage

probabilities of either infeasible and feasible confidence intervals of realized kernels are not as good as those of  $\hat{V}_{t,MA(1)}$ . Moreover,  $RK_t^F$  requires larger samples for good coverage, while density intervals of  $\hat{V}_{t,MA(1)}$  perform well for either low or high frequency returns.

The last experiment considers the performances of the estimators under dependent noise.  $RK_t^N$ ,  $RV_t$ ,  $\hat{V}_t$ ,  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  are compared. The RMSE of estimators can be found in Table 7. Again,  $RV_t$  and  $\hat{V}_t$  provide poor results if high frequency data is used. Except for one entry in the table, a version of the Bayesian estimator has the smallest RMSE in each case. The  $\hat{V}_{t,MA(1)}$  estimator is ranked the best if return frequency is 30 seconds, followed by  $\hat{V}_{t,MA(2)}$  and  $RK_t^N$ . For 10 seconds returns,  $\hat{V}_{t,MA(2)}$  provides the smallest error. Compared to  $RK_t^N$  the  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  can provide significant improvements for 30 and 10 second returns. For instance, at 30 seconds, reductions in the RMSE of 10% or more are common while at the 10 second frequency reductions in the RMSE are 25% or more.

Table 8 shows  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  have smaller bias if return frequency is one minute or higher. Table 9 shows the coverage probabilities of all the five estimators. The finite sample results of  $\hat{V}_{t,MA(2)}$  are all very close to the optimal level, no matter the data frequency.

Figure 1-3 display the histograms of the posterior mean of the number of clusters in three different settings. There are: the DPM for 5-minute SV1F returns (no noise), the DPM-MA(1) for 1-minute SV1FJ returns (independent noise) and the DPM-MA(2) for 30-second SV2F returns (dependent noise). The figures show significant pooling. For example, in the 1-minute SV1FJ return case, most of the daily variance estimates of  $V_t$  are formed by using 1 to 5 pooled groups of data, instead of 390 observations (separate groups) which is what the realized kernel uses. This level of pooling can lead to significant improvements for the Bayesian estimator.

## 5.5 Tick Time Sampling

Following Griffin & Oomen (2008), the prices simulated from DGPs illustrated in Section 5.1 are discretized to tick prices. Let  $p_{t,i}^h$  denote the observed prices and  $\omega$  represent the probability of price change.  $p_{t,i}^h = p_{t,i}$  with probability  $\omega$ , otherwise,  $p_{t,i}^h = p_{t,i-1}^h$ .  $\omega$  is set to be 0.2. Tick time returns are formed based on prices sampled every  $k^{\text{th}}$  price change. The data frequencies selected are  $k = 60$ ,  $k = 12$  and  $k = 6$ , which roughly match the frequency of 5-minute, 1-minute and 30-second data considered in previous examples.

The RMSE of estimators based on tick time sampled returns is provided in Table 10. Panel A of Table 10 compares  $RV_t$  and  $V_t$  in a no noise setting and Panel B shows the result of  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  when independent microstructure noise is present. In 10 out of 12 no-noise cases and all 12 cases with microstructure noise, the Bayesian nonparametric estimators dominate the classical counterparts in terms of RMSEs. The improvements on RMSE switching from  $RK_t^F$  to  $\hat{V}_{t,MA(1)}$  range from 6.75% to 36.70% in the cases of tick prices contaminated with noise.

Table 11 shows the RMSE of the Bayesian nonparametric estimators based on the two sampling schemes. As shown in Panel B of Table 11, the tick-time  $\hat{V}_{t,MA(1)}$  has lower RMSE in 8 out of 12 cases. However, in panel A, with no microstructure noise, calendar time sampling is uniformly better.

In summary, these simulations show the Bayesian estimate of ex-post variance to be very competitive with existing classical alternatives and under different sampling schemes.

## 6 Bayesian Nonparametric Estimates of Stock Market Variance

For each day, 5000 MCMC draws are taken after 10000 burn-in draws are discarded, to estimate posterior moments. All prior setting are the same as in the simulations.

### 6.1 IBM

We first consider estimating and forecasting volatility using a long calendar span of IBM equity returns. The 1-minute IBM price records from 1998/01/03 to 2016/02/16 were downloaded from the Kibot website<sup>10</sup>. We choose the sample starting from 2001/01/03 as the relatively small number of transactions before year 2000 yields many zero intraday returns. The days with less than 5 hours of trading are removed, which leaves 3764 days in the sample.

Log-prices are placed on a 1-minute grid using the price associated with closest time stamp that is less than or equal to the grid time. The 5-minute and 1-minute percentage log returns from 9:30 to 16:00(EST) are constructed by taking the log price difference between two close prices in time grid and scaling by 100. The overnight returns are ignored so the first intraday return is formed using the daily opening price instead of the close price in previous day. The procedure generates 293,520 5-minute returns and 1,467,848 1-minute returns.

We use a filter to remove errors and outliers caused by abnormal price records. We would like to filter out the situation in which the price jumps up or down but quickly moves back to original price range. This suggest an error in the record. If  $|r_{t,i}| + |r_{t,i+1}| > 8\sqrt{\text{var}_t(r_{t,i})}$  and  $|r_{t,i} + r_{t,i+1}| < 0.05\%$ , we replace  $r_{t,i}$  and  $r_{t,i+1}$  by  $r'_{t,i} = r'_{t,i+1} = 0.5 \times (r_{t,i} + r_{t,i+1})$ . The filter adjusts 0 and 70 ( $70/1,467,848 = 0.00477\%$ ) returns for 5-minute and 1-minute case, respectively.

From these data several version of daily  $\hat{V}_t$ ,  $RV_t$  and  $RK_t$  are computed. Daily returns are the open-to-close return and match the time interval for the variance estimates. For each of the estimators we follow exactly the methods used in the simulation section.

#### 6.1.1 Ex-post Variance Estimation

Figure 4 displays a volatility signature plot, which shows the relationship of the average volatility estimators and sampling frequency. The realized variance based on 10-minute returns serves as the unbiased benchmark because low frequency returns are less influenced by market microstructure noise. The average of the Bayes nonparametric estimator is closer to 1.0 compared with realized variance no matter the sampling frequency. The plot becomes stable after 3.9 minutes sampling frequency.

Table 12 reports summary statistics for several estimators. Overall the Bayesian and classical estimators are very close. Both the realized kernel and the moving average DPM estimators reduce the average level of daily variance and indicate the presence of significant market microstructure noise. Based on this and an analysis of the ACF of the high-frequency returns we suggest the  $\hat{V}_{t,MA(1)}$  for the 5-minute data and the  $\hat{V}_{t,MA(4)}$  for the 1-minute data in the remainder of the analysis. Comparison with the kernel estimators is found in Figures 5 and 6. Except for the extreme values they are very similar.

Interval estimates for two sub-periods are shown in Figures 7 and 8. A clear disadvantage of the kernel based confidence interval in that it includes negative values for ex-post variance.

The Bayesian version by construction does not and tends to be significantly shorter in volatile days. The results of log variance<sup>11</sup> are also provided with some differences remaining.

The degree of pooling from the Bayesian estimators is found in Figure 9 and 10. As expected, we see more groups in the higher 1-minute frequency. In this case, on average, there are about 3 to 7 distinct groups of intraday variance parameters.

### 6.1.2 Ex-post Variance Modeling and Forecasting

Does the Bayesian estimator correctly recover the time-series dynamics of volatility? To investigate this we estimate several versions of the Heterogeneous Auto-Regressive (HAR) model introduced by Corsi (2009). This is a popular model that captures the strong dependence in ex-post daily variance. For  $\hat{V}_t$  the HAR model is

$$\hat{V}_t = \beta_0 + \beta_1 \hat{V}_{t-1} + \beta_2 \hat{V}_{t-1|t-5} + \beta_3 \hat{V}_{t-1|t-22} + \epsilon_t, \quad (45)$$

where  $\hat{V}_{t-1|t-h} = \frac{1}{h} \sum_{l=1}^h \hat{V}_{t-l}$  and  $\epsilon_t$  is the error term.  $\hat{V}_{t-1}$ ,  $\hat{V}_{t-1|t-5}$  and  $\hat{V}_{t-1|t-22}$  correspond to the daily, weekly and monthly variance measures up to time  $t-1$ . Similar specifications are obtained by replacing  $\hat{V}_t$  with  $RV_t$  or  $RK_t$ .

Bollerslev et al. (2016) extend the HAR model to the HARQ model by taking the asymptotic theory of  $RV_t$  into account. The HARQ model for  $RV_t$  is given by

$$RV_t = \beta_0 + \left( \beta_1 + \beta_{1Q} RQ_{t-1}^{1/2} \right) RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \epsilon_t \quad (46)$$

The loading on  $RV_{t-1}$  is no longer a constant, but varying with measurement error, which is captured by  $RQ_{t-1}$ . The model responds more to  $RV_{t-1}$  if measurement error is low and has a lower response if error is high. Bollerslev et al. (2016) provide evidence that the HARQ model outperforms the HAR model in forecasting.<sup>12</sup>

An advantage of our Bayesian approach is that we have the full finite sample posterior distribution for  $V_t$ . In the Bayesian nonparametric framework, there is no need to estimate  $IQ_t$  with  $RQ_t$ , instead the variance, standard deviation or other features of  $V_t$  can be easily estimated using the MCMC output. Replacing  $RQ_{t-1}$  with  $\widehat{\text{var}}(V_{t-1})$ , the modified HARQ model for  $\hat{V}_t$  is defined as

$$\hat{V}_t = \beta_0 + \left( \beta_1 + \beta_{1Q} \widehat{\text{var}}(V_{t-1})^{1/2} \right) \hat{V}_{t-1} + \beta_2 \hat{V}_{t-1|t-5} + \beta_3 \hat{V}_{t-1|t-22} + \epsilon_t, \quad (47)$$

where  $\widehat{\text{var}}(V_{t-1})^{1/2}$  is an MCMC estimate of the posterior standard deviation of  $V_t$ .

Table 13 displays the OLS estimates and the  $R^2$  for several model specifications. Coefficient estimates are comparable across each class of model. Clearly the Bayesian variance estimates display the same type of time-series dynamics found in the realized kernel estimates.

Finally, out-of-sample root-mean squared forecast errors (RMSFE) of HAR and HARQ models using both classical estimators and Bayesian estimators are found in Table 14. The out-of-sample period is from 2005/01/03 to 2016/02/16 (2773 observations) and model parameters are re-estimated as new data arrives. Note, that to mimic a real-time forecast setting the prior hyperparameters  $\nu_{0,t}$  and  $s_{0,t}$  are set based on intraday data from day  $t$  and  $t-1$ .<sup>13</sup>

The first column of Table 14 reports the data frequency and the dependent variable used in the HAR/HARQ model. The second column records the data used to construct the right-hand side regressors. In this manner we consider all the possible combinations of how  $RK_t^N$  is



forecast by lags of  $RK_t^N$  or  $\hat{V}_{t,MA}$  and similarly for forecasting  $\hat{V}_{t,MA}$ . All of the specifications produce similar RMSFE. In 7 out of 8 cases the Bayesian variance measure forecasts itself and the realized kernel better.

## 6.2 SPDR S&P 500 ETF

Transaction and NBBO (National Best Bid and Offer) data for SPDR S&P 500 ETF was supplied by Tickdata. We follow the same method of Barndorff-Nielsen et al. (2011) to clean both transaction and quote datasets and form grid returns at 5-minute, 1-minute, 30-second and 10-second frequencies using transaction prices. The sample period is from July 1, 2014 to June 29, 2016 and does not include days with less than 6 trading hours. The final dataset has 498 days of intraday observations.

Table 15 displays the summary statistics of daily variance estimators of SPY returns. As the sampling frequency increases, the sample average of different variance estimators become closer to the sample variance of daily returns. Figure 11 and 12 display box plots of the daily variance estimates for the classical and Bayesian estimators for the 5-minute and 30-second data. There are several important points to make. First, both estimators recover the same general pattern of volatility in this period. Second, the Bayesian density interval is often shorter and asymmetric compare to the classical counterpart. Although there is general agreement, the high variance day of June 24 indicate some differences particularly in Figure 12. Finally, both estimates become more accurate with the higher frequency 30-second data and also make a significant downward revision to the variance estimates on June 24.

## 7 Conclusion

This paper offers a new exact finite sample approach to estimate ex-post variance using Bayesian nonparametric methods. The proposed approach benefits ex-post variance estimation in two aspects. First, the observations with similar variance levels can be pooled together to increase accuracy. Second, exact finite sample inference is available directly without relying on additional assumptions about a higher frequency DGP. Bayesian nonparametric variance estimators under no noise, heteroskedastic and serially correlated microstructure noise cases are introduced. Monte Carlo simulation results show that the proposed approach can increase the accuracy of ex-post variance estimation and provide reliable finite sample inference. Applications to real equity returns show the new estimators conform closely to the realized variance and kernel estimators in terms of average statistical properties as well as time-series characteristics. The Bayesian estimators can be used with confidence and have several benefits relative to existing methods. The Bayesian estimator can capture asymmetric density intervals, always remains positive and does not rely on the estimation of integrated quarticity.

## Notes

<sup>1</sup>For a good survey of the key concepts see Andersen & Benzoni (2008), for an in-depth treatment see Aït-Sahalia & Jacod (2014).

<sup>2</sup>Many estimators of integrated volatility calculated using high-frequency data such as realized kernels use pooling to reduce microstructure noise but are not explicitly seen as shrinkage estimators.

<sup>3</sup>Polson & Roberts (1994) is an early work that used realized variance concepts and Bayesian methods to estimate diffusion processes for stock returns.

<sup>4</sup>This infinite mixture is related to finite mixtures which have been used to approximate distributions. For example, see Kim et al. (1998).

<sup>5</sup>A MA(q) model of stationary and ergodic microstructure noise would not allow for staleness (Bandi et al. 2017) or flat trading (Phillips & Yu 2007).

<sup>6</sup>Restrictions on MA coefficients: all the roots of  $1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q = 0$  are outside of the unit circle.

<sup>7</sup>If  $\tilde{r}_t = \theta_1 \eta_{t-1} + \dots + \theta_q \eta_{t-q+1} + \eta_t$ , then under their assumptions the bias corrected estimate of ex-post variance is  $RV_{MAq} = (1 + \theta_1 + \dots + \theta_q)^2 \sum_{i=1}^{n_t} \hat{\eta}_i^2$ , where  $\hat{\eta}_i$  denotes a fitted residual.

<sup>8</sup>The function  $s\text{-exp}$  is defined as  $s\text{-exp}(x) = \exp(x)$  if  $x \leq x_0$  and  $s\text{-exp}(x) = \frac{\exp(x_0)}{\sqrt{x_0}} \sqrt{x_0 - x_0^2 + x^2}$  if  $x > x_0$ , with  $x_0 = \log(1.5)$ .

<sup>9</sup>The blocked realized variance is calculated as follows.  $RV_t^{block} = M \Delta t \sum_i \frac{M \Gamma(\frac{M-1}{2})}{\Gamma(\frac{M+1}{2})} \sigma_{\tau_i}^2$  where  $\sigma_{\tau_i}^2 = \frac{1}{\Delta t(M-1)} \sum_{t_j \in (\tau_i, \tau_{i+1})} (r_{t_j} - \bar{r}_{\tau_i})^2$ ,  $\bar{r}_{\tau_i} = \frac{1}{M} \sum_{t_j \in (\tau_i, \tau_{i+1})} r_{t_j}$  and  $M$  is the block size.

<sup>10</sup><http://www.kibot.com>

<sup>11</sup>95% confidence intervals using  $\log(RV_t)$ ,  $\log(RK_t^F)$  and  $\log(RK_t^N)$  are based on the asymptotic distributions in Barndorff-Nielsen & Shephard (2002), Barndorff-Nielsen et al. (2008) and Barndorff-Nielsen et al. (2011).

<sup>12</sup>A drawback of this specification is that it is possible for the coefficient on  $RV_{t-1}$  to be negative and produce a negative forecast for next period's variance. To avoid this when  $\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2} < 0$  it is set to 0.

<sup>13</sup>Data from day  $t+1$  would not be available in a real-time scenario. Using only data from day  $t$  to set  $\nu_{0,t}$  and  $s_{0,t}$  gives very similar results.

## References

- Aït-Sahalia, Y. & Jacod, J. (2014), *High-Frequency Financial Econometrics*, Princeton University Press.
- Aït-Sahalia, Y. & Mancini, L. (2008), 'Out of sample forecasts of quadratic variation', *Journal of Econometrics* **147**(1), 17 – 33.
- Aït-Sahalia, Y., Mykland, P. A. & Zhang, L. (2011), 'Ultra high frequency volatility estimation with dependent microstructure noise', *Journal of Econometrics* **160**(1), 160–175.
- Andersen, T. & Bollerslev, T. (1998), 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**(4), 885–905.
- Andersen, T. G. & Benzoni, L. (2008), Realized volatility, Working Paper Series WP-08-14, Federal Reserve Bank of Chicago.
- Andersen, T. G., Bollerslev, T., Diebold, F. X. & Ebens, H. (2001), 'The distribution of realized stock return volatility', *Journal of Financial Economics* **61**(1), 43–76.
- Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The distribution of realized exchange rate volatility', *Journal of the American Statistical Association* **96**, 42–55.

- Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), ‘Modeling and forecasting realized volatility’, *Econometrica* **71**(2), 579–625.
- Andersen, T. G., Bollerslev, T. & Meddahi, N. (2011), ‘Realized volatility forecasting and market microstructure noise’, *Journal of Econometrics* **160**(1), 220 – 234.
- Bandi, F. M., Pirino, D. & Reno, R. (2017), ‘EXcess Idle Time’, *Econometrica* **85**(6), 1793–1846.
- Bandi, F. M. & Russell, J. R. (2008), ‘Microstructure noise, realized variance, and optimal sampling’, *The Review of Economic Studies* **75**(2), 339–369.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2008), ‘Designing realized kernels to measure the ex post variation of equity prices in the presence of noise’, *Econometrica* **76**(6), 1481–1536.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2009), ‘Realized kernels in practice: trades and quotes’, *Econometrics Journal* **12**(3), C1–C32.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2011), ‘Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading’, *Journal of Econometrics* **162**(2), 149–169.
- Barndorff-Nielsen, O. E. & Shephard, N. (2002), ‘Estimating quadratic variation using realized variance’, *Journal of Applied Econometrics* **17**, 457–477.
- Barndorff-Nielsen, O. E. & Shephard, N. (2006), ‘Econometrics of testing for jumps in financial economics using bipower variation’, *Journal of Financial Econometrics* **4**(1), 1–30.
- Bollen, B. & Inder, B. (2002), ‘Estimating daily volatility in financial markets utilizing intraday data’, *Journal of Empirical Finance* **9**(5), 551–562.
- Bollerslev, T., Patton, A. & Quaedvlieg, R. (2016), ‘Exploiting the errors: A simple approach for improved volatility forecasting’, *Journal of Econometrics* pp. 1–18.
- Brown, L. D. & Zhao, L. H. (2012), ‘A geometrical explanation of Stein shrinkage’, *Statist. Sci.* **27**(1), 24–30.
- Chernov, M., Ronald Gallant, A., Ghysels, E. & Tauchen, G. (2003), ‘Alternative models for stock price dynamics’, *Journal of Econometrics* **116**(1-2), 225–257.
- Corradi, V., Distaso, W. & Swanson, N. R. (2009), ‘Predictive density estimators for daily volatility based on the use of realized measures’, *Journal of Econometrics* **150**(2), 119 – 138.
- Corsi, F. (2009), ‘A simple approximate long-memory model of realized volatility’, *Journal of Financial Econometrics* **7**(2), 174–196.
- Escobar, M. D. & West, M. (1994), ‘Bayesian density estimation and inference using mixtures’, *Journal of the American Statistical Association* **90**, 577–588.

- Ferguson, T. S. (1973), ‘A Bayesian analysis of some nonparametric problems’, *The Annals of Statistics* **1**(2), 209–230.
- Goncalves, S. & Meddahi, N. (2009), ‘Bootstrapping realized volatility’, *Econometrica* **77**(1), 283–306.
- Griffin, J. E. & Oomen, R. C. A. (2008), ‘Sampling returns for realized variance calculations: Tick time or transaction time?’, *Econometric Reviews* **27**(1-3), 230–253.
- Hansen, P., Large, J. & Lunde, A. (2008), ‘Moving average-based estimators of integrated variance’, *Econometric Reviews* **27**(1-3), 79–111.
- Hansen, P. R. & Lunde, A. (2006), ‘Realized variance and market microstructure noise’, *Journal of Business & Economic Statistics* **24**, 127–161.
- Huang, X. & Tauchen, G. (2005), ‘The relative contribution of jumps to total price variance’, *Journal of Financial Econometrics* **3**(4), 456–499.
- Kalli, M., Griffin, J. E. & Walker, S. G. (2011), ‘Slice sampling mixture models’, *Statistics and Computing* **21**(1), 93–105.
- Kim, S., Shephard, N. & Chib, S. (1998), ‘Stochastic volatility: Likelihood inference and comparison with ARCH models’, *The Review of Economic Studies* **65**(3), 361–393.
- Maheu, J. M. & McCurdy, T. H. (2002), ‘Nonlinear features of realized FX volatility’, *The Review of Economics and Statistics* **84**(4), 668–681.
- Mykland, P. A. & Zhang, L. (2009), ‘Inference for continuous semimartingales observed at high frequency’, *Econometrica* **77**(5), 1403–1445.
- Phillips, P. C. B. & Yu, J. (2007), Information loss in volatility measurement with flat price trading. Cowles Foundation Discussion Paper 1598.
- Polson, N. G. & Roberts, G. O. (1994), ‘Bayes factors for discrete observations from diffusion processes’, *Biometrika* **81**(1), 11–26.
- Sethuraman, J. (1994), ‘A constructive definition of Dirichlet priors’, *Statistica Sinica* pp. 639–650.
- Zhang, L., Mykland, P. A. & Ait-Sahalia, Y. (2005), ‘A tale of two time scales: Determining integrated volatility with noisy high-frequency data’, *Journal of the American Statistical Association* **100**, 1394–1411.
- Zhou, B. (1996), ‘High frequency data and volatility in foreign exchange rates’, *Journal of Business and Economic Statistics* **14**(1), 45–52.

## 8 Appendix

### 8.1 Existing Ex-post Volatility Estimation

#### 8.1.1 Realized Variance

Let  $r_{t,i}$  denotes the  $i^{\text{th}}$  intraday return on day  $t$ ,  $i = 1, \dots, n_t$ , where  $n_t$  is the number of intraday returns on day  $t$ . Realized variance is defined as

$$RV_t = \sum_{i=1}^{n_t} r_{t,i}^2, \quad (48)$$

and  $RV_t \xrightarrow{p} IV_t$ , as  $n_t \rightarrow \infty$  (Andersen, Bollerslev, Diebold & Labys 2001).

Barndorff-Nielsen & Shephard (2002) derive the asymptotic distribution of  $RV_t$  as

$$\sqrt{n_t} \frac{1}{\sqrt{2IQ_t}} (RV_t - IV_t) \xrightarrow{d} N(0, 1), \quad \text{as } n_t \rightarrow \infty, \quad (49)$$

where  $IQ_t$  stands for the integrated quarticity, which can be estimated by realized quarticity ( $RQ_t$ ) defined as

$$RQ_t = \frac{n_t}{3} \sum_{i=1}^{n_t} r_{t,i}^4 \xrightarrow{p} IQ_t, \quad \text{as } n_t \rightarrow \infty. \quad (50)$$

#### 8.1.2 Flat-top Realized Kernel

Barndorff-Nielsen et al. (2008) introduced the flat-top realized kernel ( $RK_t^F$ ), which is the optimal estimator if the microstructure error is a white noise process<sup>14</sup>.

$$RK_t^F = \sum_{i=1}^{n_t} \tilde{r}_{t,i}^2 + \sum_{h=1}^H k \left( \frac{h-1}{H} \right) (\gamma_{-h} + \gamma_h), \quad \gamma_h = \sum_{i=1}^{n_t} \tilde{r}_{t,i} \tilde{r}_{t,i-h}, \quad (51)$$

where  $H$  is the bandwidth,  $k(x)$  is a kernel weight function.

The preferred kernel function is the second order Tukey-Hanning kernel<sup>15</sup> and the preferred bandwidth is  $H^* = c\xi\sqrt{n_t}$ , where  $\xi^2 = \omega^2/\sqrt{IQ_t}$  denotes the noise-to-signal ratio.  $\omega^2$  stands for the variance of microstructure noise and can be estimated by  $RV_t/(2n_t)$  by Bandi & Russell (2008).  $RV_t$  based on 10-minute returns is less sensitive to microstructure noise and can be used as a proxy of  $\sqrt{IQ_t}$ .  $c = 5.74$  given Tukey-Hanning kernel of order 2.

Given the Tukey-Hanning kernel and  $H^* = c\xi\sqrt{n_t}$ , Barndorff-Nielsen et al. (2008) show that the asymptotic distribution of  $RK_t^F$  is

$$n_t^{1/4} (RK_t^F - IV_t) \xrightarrow{d} MN \left\{ 0, 4IQ_t^{3/4} \omega \left( ck_{\bullet}^{0,0} + 2c^{-1}k_{\bullet}^{1,1} \frac{IV_t}{\sqrt{IQ_t}} + c^{-3}k_{\bullet}^{2,2} \right) \right\}, \quad (52)$$

where MN is mixture of normal distribution,  $k_{\bullet}^{0,0} = 0.219$ ,  $k_{\bullet}^{1,1} = 1.71$  and  $k_{\bullet}^{2,2} = 41.7$  for second order Tukey-Hanning kernel.

Even though  $\omega^2$  can be estimated using  $RV_t/(2n_t)$ , a better and less biased estimator suggested by Barndorff-Nielsen et al. (2008) is

$$\tilde{\omega}^2 = \exp \left[ \log(\hat{\omega}^2) - RK_t/RV_t \right]. \quad (53)$$

The estimation of  $IQ_t$  is more sensitive to the microstructure noise. The tri-power quarticity ( $TPQ_t$ ) developed by Barndorff-Nielsen & Shephard (2006) can be used to estimate  $IQ_t$ ,

$$TPQ_t = n_t \mu_{4/3}^{-3} \sum_{i=1}^{n_t-2} |\tilde{r}_{t,i}|^{4/3} |\tilde{r}_{t,i+1}|^{4/3} |\tilde{r}_{t,i+2}|^{4/3}, \quad (54)$$

where  $\mu_{4/3} = 2^{2/3} \Gamma(7/6) / \Gamma(1/2)$ . Replacing  $IV_t$ ,  $\omega^2$  and  $IQ_t$  with  $RK_t^F$ ,  $\tilde{\omega}^2$  and  $TPQ_t$  in equation (52), the asymptotic variance of  $RK_t^F$  can be calculated.

### 8.1.3 Non-negative Realized Kernel

The flat-top realized kernel discussed in previous subsection is based on the assumption that the error term is white noise. However, the white noise assumption is restrictive and the error term can be serial dependent or dependent on returns in reality. Another drawback of the  $RK_t^F$  is that it may provide negative volatility estimates, albeit very rarely. Barndorff-Nielsen et al. (2011) further introduced the non-negative realized kernel ( $RK_t^N$ ) which is more robust to these assumptions of error term and is calculated as

$$RK_t^N = \sum_{h=-H}^H k\left(\frac{h}{H+1}\right) \gamma_h, \quad \gamma_h = \sum_{i=|h|+1}^{n_t} \tilde{r}_{t,i} \tilde{r}_{t,i-|h|}. \quad (55)$$

The optimal choice of  $H$  is  $H^* = c \xi^{4/5} n_t^{3/5}$  and the preferred kernel weight function is the Parzen kernel<sup>16</sup>, which implies  $c = 3.5134$ .  $\xi^2$  can be estimated using the same method as in the calculation of  $RK_t^F$ .

Barndorff-Nielsen et al. (2011) show the asymptotic distribution of  $RK_t^N$  based on  $H^* = c \xi^{4/5} n_t^{3/5}$  is given by

$$n_t^{1/5} (RK_t^N - IV_t) \xrightarrow{d} \text{MN}(\kappa, 4\kappa^2), \quad (56)$$

where  $\kappa = \kappa_0 (IQ_t \omega)^{2/5}$ ,  $\kappa_0 = 0.97$  for Parzen kernel function,  $\omega$  and  $IQ_t$  can be estimated using equation (53) and (54).

## 8.2 Adjustment to DPM-MA(1) Estimator

Let  $p_{t,i}$  denotes the latent intraday price and  $\epsilon_{t,i}$  is the microstructure noise which is independently distributed and heteroskedastic. The observed intraday price  $\tilde{p}_{t,i}$  is

$$\tilde{p}_{t,i} = p_{t,i} + \epsilon_{t,i}, \quad E(\epsilon_{t,i}) = 0 \text{ and } \text{var}(\epsilon_{t,i}) = \omega_{t,i}^2. \quad (57)$$

The log return process is constructed as follows,

$$\tilde{r}_{t,i} = \tilde{p}_{t,i} - \tilde{p}_{t,i-1} = p_{t,i} - p_{t,i-1} + \epsilon_{t,i} - \epsilon_{t,i-1} = r_{t,i} + \epsilon_{t,i} - \epsilon_{t,i-1}, \quad (58)$$

where  $\tilde{r}_{t,i}$  and  $r_{t,i}$  are the observed return and pure return. The variance and first autocovariance of  $\{r_{t,i}\}_{i=1}^{n_t}$  are

$$\text{var}(\tilde{r}_{t,i}) = \sigma_{t,i}^2 + \omega_{t,i}^2 + \omega_{t,i-1}^2, \quad (59)$$

$$\text{cov}(\tilde{r}_{t,i}, \tilde{r}_{t,i-1}) = -\omega_{t,i-1}^2. \quad (60)$$

Consider the following heteroskedastic MA(1) model for the observed  $\tilde{r}_{t,i}$ ,

$$\tilde{r}_{t,i} = \mu_t + \theta_t \eta_{t,i-1} + \eta_{t,i}, \quad \eta_{t,i} \sim N(0, \delta_{t,i}^2), \quad (61)$$

which will be used to recover an estimate of ex-post variance for the pure return process,  $V_t = \sum_{i=1}^{n_t} \sigma_{t,i}^2$ . The corresponding moments of this process are

$$\text{var}(\tilde{r}_{t,i}) = \theta_t^2 \delta_{t,i-1}^2 + \delta_{t,i}^2, \quad (62)$$

$$\text{cov}(\tilde{r}_{t,i}, \tilde{r}_{t,i-1}) = \theta_t \delta_{t,i-1}^2. \quad (63)$$

Equating (59) and (62), we have

$$\sigma_{t,i}^2 + \omega_{t,i}^2 + \omega_{t,i-1}^2 = \theta_t^2 \delta_{t,i-1}^2 + \delta_{t,i}^2 \quad (64)$$

Equating (60) and (63), we have

$$-\omega_{t,i-1}^2 = \theta_t \delta_{t,i-1}^2 \quad \text{and} \quad -\omega_{t,i}^2 = \theta_t \delta_{t,i}^2. \quad (65)$$

Based on the result in (65), the summation of  $\delta_{t,i}^2$ , over  $i = 1, \dots, n_t$ , equals

$$\sum_{i=1}^{n_t} \delta_{t,i}^2 = -\frac{1}{\theta_t} \sum_{i=1}^{n_t} \omega_{t,i}^2. \quad (66)$$

Plugging both terms in (65) into (64), yields

$$\sigma_{t,i}^2 + \omega_{t,i}^2 + \omega_{t,i-1}^2 = -\theta_t \omega_{t,i-1}^2 - \frac{\omega_{t,i}^2}{\theta_t} \quad (67)$$

$$\sigma_{t,i}^2 + \left(1 + \frac{1}{\theta_t}\right) \omega_{t,i}^2 + (1 + \theta_t) \omega_{t,i-1}^2 = 0. \quad (68)$$

Using the results in (68), the summation of  $\sigma_{t,i}^2$ , over  $i = 1, \dots, n_t$ , equals

$$\sum_{i=1}^{n_t} \sigma_{t,i}^2 + \left(1 + \frac{1}{\theta_t}\right) \sum_{i=1}^{n_t} \omega_{t,i}^2 + (1 + \theta_t) \sum_{i=1}^{n_t} \omega_{t,i-1}^2 = 0 \quad (69)$$

$$V_t = -\left(1 + \frac{1}{\theta_t}\right) \sum_{i=1}^{n_t} \omega_{t,i}^2 - (1 + \theta_t) \sum_{i=1}^{n_t} \omega_{t,i-1}^2. \quad (70)$$

The ratio between (66) and (70) is

$$\frac{V_t}{\sum_{i=1}^{n_t} \delta_{t,i}^2} = \frac{-\left(1 + \frac{1}{\theta_t}\right) \sum_{i=1}^{n_t} \omega_{t,i}^2 - (1 + \theta_t) \sum_{i=1}^{n_t} \omega_{t,i-1}^2}{-\frac{1}{\theta_t} \sum_{i=1}^{n_t} \omega_{t,i}^2} \quad (71)$$

$$= \frac{(1 + \theta_t)^2 \sum_{i=1}^{n_t-1} \omega_{t,i}^2 + (1 + \theta_t) \omega_{t,n_t}^2 + (\theta_t + \theta_t^2) \omega_{t,0}^2}{\sum_{i=1}^{n_t-1} \omega_{t,i}^2 + \omega_{t,n_t}^2} \quad (72)$$

$$= (1 + \theta_t)^2, \quad \text{if } \omega_{t,n_t} = \omega_{t,0}. \quad (73)$$

Finally, we have

$$(1 + \theta_t)^2 \sum_{i=1}^{n_t} \delta_{t,i}^2 = V_t, \quad \text{if } \omega_{t,n_t} = \omega_{t,0}. \quad (74)$$

### 8.3 Adjustment to DPM-MA(2) Estimator

If the observed intraday price  $\tilde{p}_{t,i}$  is

$$\tilde{p}_{t,i} = p_{t,i} + \epsilon_{t,i} - \rho\epsilon_{t,i-1}, \quad E(\epsilon_{t,i}) = 0 \text{ and } \text{var}(\epsilon_{t,i}) = \omega_{t,i}^2. \quad (75)$$

Then log return process is constructed as follows.

$$\begin{aligned} \tilde{r}_{t,i} &= \tilde{p}_{t,i} - \tilde{p}_{t,i-1} \\ &= r_{t,i} + \epsilon_{t,i} - (1 + \rho)\epsilon_{t,i-1} + \rho\epsilon_{t,i-2}. \end{aligned} \quad (76)$$

Using the following heteroskedastic MA(2) model for  $\tilde{r}_{t,i}$ ,

$$\tilde{r}_{t,i} = \mu_t + \theta_{1t}\eta_{t,i-1} + \theta_{2t}\eta_{t,i-2} + \eta_{t,i}, \quad \eta_{t,i} \sim N(0, \delta_{t,i}^2) \quad (77)$$

it can be shown the adjustment term is

$$(1 + \theta_{1t} + \theta_{2t})^2 \sum_{i=1}^{n_t} \delta_{t,i}^2 = V_t, \quad \text{if } \omega_{t,n_t-1} = \omega_{t,0} \text{ and } \omega_{t,n_t} = \omega_{t,-1}. \quad (78)$$

Similar results hold for higher order MA models.

### 8.4 Proof of Theorem 1

We will show that  $E[V_t | \{r_{t,i}\}_{i=1}^{n_t}]$  can be expressed as realized volatility,  $RV_t$ , plus a bias term and that this bias converges to zero. It is useful to introduce the allocation variables used in MCMC sampler in Section 2.3 to derive a suitable expression for  $E[V_t | \{r_{t,i}\}_{i=1}^{n_t}]$  by using

$$E[V_t | \{r_{t,i}\}_{i=1}^{n_t}] = E \left[ \sum_{i=1}^{n_t} \sigma_{t,i}^2 \middle| \{r_{t,i}\}_{i=1}^{n_t} \right] = E_{\{s_{t,i}\}_{i=1}^{n_t} | \{r_{t,i}\}_{i=1}^{n_t}} E \left[ \sum_{i=1}^{n_t} \sigma_{t,i}^2 \middle| \{r_{t,i}\}_{i=1}^{n_t}, \{s_{t,i}\}_{i=1}^{n_t} \right].$$

Conditional on the allocation variables  $\{s_{t,i}\}_{i=1}^{n_t}$ , we can write

$$\sigma_t^2 = \sum_{i=1}^{n_t} \sigma_{t,i}^2 = \sum_{k=1}^{K_t} m_{t,k} \psi_{t,k},$$

where  $K_t$  is the number of distinct values of  $\{s_{t,i}\}_{i=1}^{n_t}$ ,  $m_{t,k} = \sum_{i=1}^{n_t} \mathbb{1}(s_{t,i} = k)$  and  $\psi_{t,k} \sim IG(\nu_{0,t} + m_{t,k}/2, s_{0,t} + \sum_{s_{t,i}=k} r_{t,i}^2/2)$ . It follows that

$$\begin{aligned} E \left[ \sum_{i=1}^{n_t} \sigma_{t,i}^2 \middle| \{r_{t,i}\}_{i=1}^{n_t}, \{s_{t,i}\}_{i=1}^{n_t} \right] &= \sum_{k=1}^{K_t} m_{t,k} \frac{s_{0,t} + \sum_{s_{t,i}=k} r_{t,i}^2/2}{\nu_{0,t} + m_{t,k}/2 - 1} \\ &= s_{0,t} \sum_{k=1}^{K_t} \frac{m_{t,k}}{\nu_{0,t} + m_{t,k}/2 - 1} + \sum_{k=1}^{K_t} \left( 1 - \frac{\nu_{0,t} - 1}{\nu_{0,t} + m_{t,k}/2 - 1} \right) \sum_{s_{t,i}=k} r_{t,i}^2 \\ &= \sum_{i=1}^{n_t} r_{t,i}^2 + s_{0,t} \sum_{k=1}^{K_t} \frac{m_{t,k}}{\nu_{0,t} + m_{t,k}/2 - 1} - \sum_{k=1}^{K_t} \frac{\nu_{0,t} - 1}{\nu_{0,t} + m_{t,k}/2 - 1} \sum_{s_{t,i}=k} r_{t,i}^2 \\ &= RV_t + B(s_{t,1:n_t}) \end{aligned}$$



where

$$\begin{aligned} B(\{s_{t,i}\}_{i=1}^{n_t}) &= \sum_{k=1}^{K_t} \frac{[s_{0,t}m_{t,k} - (\nu_{0,t} - 1) \sum_{s_{t,i}=k} r_{t,i}^2]}{\nu_{0,t} + m_{t,k}/2 - 1} \\ &= \sum_{k=1}^{K_t} \frac{m_{t,k}}{\nu_{0,t} + m_{t,k}/2 - 1} \left[ s_{0,t} - (\nu_{0,t} - 1) \frac{1}{m_{t,k}} \sum_{s_{t,i}=k} r_{t,i}^2 \right]. \end{aligned}$$

The posterior mean can be expressed as

$$E[V_t | \{r_{t,i}\}_{i=1}^{n_t}] = RV_t + E_{\{s_{t,i}\}_{i=1}^{n_t} | \{r_{t,i}\}_{i=1}^{n_t}} [B(\{s_{t,i}\}_{i=1}^{n_t})]. \quad (79)$$

To show consistency, we use the result that  $RV_t$  is a consistent estimator of quadratic variation under our conditions (Andersen et al. 2003). To see that the bias term  $E_{\{s_{t,i}\}_{i=1}^{n_t} | \{r_{t,i}\}_{i=1}^{n_t}} [B(\{s_{t,i}\}_{i=1}^{n_t})]$  converges in probability to zero, notice that

$$\begin{aligned} E_{\{s_{t,i}\}_{i=1}^{n_t} | \{r_{t,i}\}_{i=1}^{n_t}} [B(\{s_{t,i}\}_{i=1}^{n_t})] &= \sum_{k=1}^{K_t} \frac{m_{t,k}}{\nu_{0,t} + m_{t,k}/2 - 1} \left[ s_{0,t} - (\nu_{0,t} - 1) \frac{1}{m_{t,k}} \sum_{s_{t,i}=k} \sigma_{t,i}^2 \right] \\ &< 2 \left[ K_t s_{0,t} - (\nu_{0,t} - 1) \sum_{k=1}^{K_t} \frac{1}{m_{t,k}} \sum_{s_{t,i}=k} \sigma_{t,i}^2 \right]. \end{aligned}$$

Clearly, this converges in probability to zero if  $K_t s_{0,t} \rightarrow 0$  as  $\frac{1}{m_{t,k}} \sum_{s_{t,i}=k} \sigma_{t,i}^2 \rightarrow 0$  in probability for all  $s_{t,1:n_t}$  (since  $\sup(\tau_{t,j+1} - \tau_{t,j}) \rightarrow 0$ ). We know that  $E[K_t] \approx M \log(1 + \frac{n}{M})$  and the results follows from the assumption that  $s_{0,n_t} = O(n_t^{-\alpha})$  for  $\alpha > 0$ .

Table 1: Prior Specifications of Models

Model	$\mu_t$	$\sigma_{t,i}^2$	$\Theta_t$	$\alpha_t$
DPM	$N(0, v^2)$	$IG(v_{0,t}, s_{0,t})$	-	Gamma(2, 8)
DPM-MA(q)	$N(0, v^2)$	$IG(v_{0,t}, s_{0,t})$	$N(\mathbf{0}, I)\mathbb{1}_{\{\Theta_t\}}$	Gamma(2, 8)

<sup>1.</sup>  $v_{0,t}$  and  $s_{0,t}$  are calculated using equation (12).

<sup>2.</sup>  $\mathbb{1}_{\{\Theta_t\}}$  denotes the invertibility condition for the MA(q) model.

<sup>3.</sup>  $v^2 = \zeta^2/n_t$  where  $\zeta^2 = 0.01$  and  $n_t$  is the number of intraday returns.

Table 2: RMSE of  $RV_t$ , blocked  $RV_t$  and  $\hat{V}_t$  (No Microstructure Noise Case)

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	0.12352	0.21226	0.21471	0.45601
	$RV_{t,M=26}^{block}$	0.12485	0.21588	0.21680	0.44368
	$\hat{V}_t$	<b>0.11866***</b>	<b>0.20116***</b>	<b>0.20659***</b>	<b>0.43095**</b>
1-minute	$RV_t$	0.05368	0.09283	<b>0.09771</b>	0.23296
	$RV_{t,M=78}^{block}$	0.05397	0.09335	0.09875	0.23120
	$\hat{V}_t$	<b>0.05323***</b>	<b>0.09190***</b>	0.10051	<b>0.22802*</b>
30-second	$RV_t$	0.03886	0.06530	<b>0.06741*</b>	0.14178
	$RV_{t,M=130}^{block}$	0.03906	0.06539	0.06771	0.14184
	$\hat{V}_t$	<b>0.03867***</b>	<b>0.06495***</b>	0.07276	<b>0.13970</b>
10-second	$RV_t$	0.02177	0.03601	<b>0.03662*</b>	<b>0.09535</b>
	$RV_{t,M=260}^{block}$	0.02176	0.03601	0.03677	0.09587
	$\hat{V}_t$	<b>0.02171**</b>	<b>0.03589***</b>	0.04722	0.09596

<sup>1.</sup> This table reports the root mean squared error (RMSE) of estimating 5000 daily ex-post variances using  $RV_t$ , blocked  $RV_t$  and Bayesian nonparametric estimator  $\hat{V}_t$  under different frequencies and DGPs. Microstructure noise is not considered.

<sup>2.</sup> A paired t-test is used to test whether the difference in the mean of  $(RV_t - V_t)^2$  and  $(\hat{V}_t - V_t)^2$  is equal to zero. The symbol \* indicates the levels of significance. \*: p-value < 0.05, \*\*: p-value < 0.01, \*\*\*: p-value < 0.001.

Table 3: Prior Robustness Check

Estimator	Prior of $\mu_t$	Prior of $\alpha_t$	Prior of $\sigma_{t,i}^2$	RMSE
<i>Panel A: 5-minute return</i>				
$RV_t$	-	-	-	0.45601
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.43095
	$N(0, \frac{0.05}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.42919
$\hat{V}_t$	$N(0, \frac{0.002}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.42953
	$N(0, \frac{0.01}{n_t})$	Gamma(2, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.42975
	$N(0, \frac{0.01}{n_t})$	Gamma(8, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.42782
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.43199
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.43199
<i>Panel B: 1-minute return</i>				
$RV_t$	-	-	-	0.23296
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.22802
	$N(0, \frac{0.05}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.22739
$\hat{V}_t$	$N(0, \frac{0.002}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.22686
	$N(0, \frac{0.01}{n_t})$	Gamma(2, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.22691
	$N(0, \frac{0.01}{n_t})$	Gamma(8, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.22613
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.22695
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.22695
<i>Panel C: 30-second return</i>				
$RV_t$	-	-	-	0.14178
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.13970
	$N(0, \frac{0.05}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.14059
$\hat{V}_t$	$N(0, \frac{0.002}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.14012
	$N(0, \frac{0.01}{n_t})$	Gamma(2, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.14096
	$N(0, \frac{0.01}{n_t})$	Gamma(8, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.14003
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.14029
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.14029
<i>Panel D: 10-second return</i>				
$RV_t$	-	-	-	0.09535
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.09596
	$N(0, \frac{0.05}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.09631
$\hat{V}_t$	$N(0, \frac{0.002}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.09546
	$N(0, \frac{0.01}{n_t})$	Gamma(2, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.09610
	$N(0, \frac{0.01}{n_t})$	Gamma(8, 8)	$IG(v_{0,t}^{3d}, s_{0,t}^{3d})$	0.09644
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.09528
	$N(0, \frac{0.01}{n_t})$	Gamma(4, 8)	$IG(v_{0,t}^{1d}, s_{0,t}^{1d})$	0.09528

This table reports the root mean squared error (RMSE) of estimating 5000 daily ex-post variances using  $RV_t$  and Bayes nonparametric estimator  $\hat{V}_t$  under different priors.  $v_{0,t}^{3d}$  denotes 3 days used to calibrate the prior parameter while  $v_{0,t}^{1d}$  denotes one day of data used. The data are generated from SV2F. Microstructure noise is not considered.

Table 4: RMSE of  $RV_t$  and  $\hat{V}_t$  (Irregular-spaced Case)

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
$\lambda = 300$	$RV_t$	0.15890	0.34273	0.32518	0.70050
	$\hat{V}_t$	<b>0.14724</b>	<b>0.31481</b>	<b>0.30301</b>	<b>0.67717</b>
$\lambda = 60$	$RV_t$	0.14903	0.15044	0.14741	0.63304
	$\hat{V}_t$	<b>0.14297</b>	<b>0.14611</b>	<b>0.14524</b>	<b>0.57660</b>
$\lambda = 30$	$RV_t$	0.06989	0.09738	<b>0.10187</b>	0.27621
	$\hat{V}_t$	<b>0.06827</b>	<b>0.09604</b>	0.10355	<b>0.26985</b>
$\lambda = 10$	$RV_t$	0.03399	0.06432	<b>0.06575</b>	0.23201
	$\hat{V}_t$	<b>0.03380</b>	<b>0.06392</b>	0.06851	<b>0.23070</b>

We follow Barndorff-Nielsen et al. (2011) to simulate irregular-spaced prices. The arrival times of observations are simulated from a Poisson process. The parameter  $\lambda$  in Poisson process governs the trading frequency of simulated data. For example,  $\lambda = 30$  means the transactions arrive every 30 seconds on average.

Table 5: RMSE of  $RV_t$ ,  $RK_t^F$ ,  $\hat{V}_t$  and  $\hat{V}_{t,MA(1)}$  (Independent Microstructure Error Case)

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	0.16003	0.29182	0.30651	0.47783
	$RK_t^F$	0.22988	0.42318	0.43993	0.84100
	$\hat{V}_t$	<b>0.15640</b>	<b>0.28464</b>	<b>0.29858</b>	<b>0.46117</b>
	$\hat{V}_{t,MA(1)}$	0.21636	0.38776	0.40729	0.74828
1-minute	$RV_t$	0.48607	0.85374	0.94598	0.63983
	$RK_t^F$	0.11157	0.20184	0.20822	0.46655
	$\hat{V}_t$	0.48735	0.85547	0.94689	0.63808
	$\hat{V}_{t,MA(1)}$	<b>0.10592</b>	<b>0.18787</b>	<b>0.19539</b>	<b>0.41176</b>
30-second	$RV_t$	0.95855	1.69544	1.87445	1.20299
	$RK_t^F$	0.08483	0.15200	0.15743	<b>0.27201</b>
	$\hat{V}_t$	0.96016	1.69798	1.87569	1.20332
	$\hat{V}_{t,MA(1)}$	<b>0.07906</b>	<b>0.14017</b>	<b>0.15232</b>	0.27595
10-second	$RV_t$	2.86639	5.06382	5.60527	3.57263
	$RK_t^F$	0.05575	0.10097	0.10683	<b>0.16989</b>
	$\hat{V}_t$	2.86858	5.06745	5.60757	3.57263
	$\hat{V}_{t,MA(1)}$	<b>0.05387</b>	<b>0.09621</b>	<b>0.10555</b>	0.20857

This table reports the root mean squared error (RMSE) of estimating 5000 daily ex-post variances using  $RV_t$ ,  $RK_t^F$  and Bayesian nonparametric estimators  $\hat{V}_t$  and  $\hat{V}_{t,MA(1)}$  based on returns at different frequencies and simulated from 4 DGPs. The price is contaminated with white noise.

Table 6: Coverage Probability (Independent Microstructure Error Case)

Data Freq.	Interval Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	87.60%	85.00%	84.42%	21.56%
	$RK_t^F$ - Infeasible	87.84%	87.66%	87.94%	93.48%
	$RK_t^F$ - Feasible	84.28%	96.20%	83.68%	97.72%
	$\hat{V}_t$	81.84%	78.50%	77.10%	18.12%
	$\hat{V}_{t,MA(1)}$	94.02%	94.40%	94.14%	89.74%
1-minute	$RV_t$	0.46%	0.82%	0.78%	5.24%
	$RK_t^F$ - Infeasible	88.50%	89.78%	89.02%	93.32%
	$RK_t^F$ - Feasible	99.30%	97.76%	95.26%	97.86%
	$\hat{V}_t$	0.42%	0.72%	0.56%	4.48%
	$\hat{V}_{t,MA(1)}$	95.06%	95.18%	94.66%	86.60%
30-second	$RV_t$	0.00%	0.00%	0.02%	1.72%
	$RK_t^F$ - Infeasible	89.80%	90.46%	90.74%	92.80%
	$RK_t^F$ - Feasible	77.44%	99.48%	99.52%	97.94%
	$\hat{V}_t$	0.00%	0.00%	0.00%	1.54%
	$\hat{V}_{t,MA(1)}$	95.00%	95.18%	94.84%	85.94%
10-second	$RV_t$	0.00%	0.00%	0.00%	0.04%
	$RK_t^F$ - Infeasible	92.08%	92.68%	92.90%	92.10%
	$RK_t^F$ - Feasible	99.98%	99.98%	99.98%	98.62%
	$\hat{V}_t$	0.00%	0.00%	0.00%	0.04%
	$\hat{V}_{t,MA(1)}$	94.92%	95.34%	95.32%	82.24%

This table reports the coverage probabilities of 95% confidence intervals using  $RV_t$ ,  $RK_t^F$  and 0.95 density intervals using  $\hat{V}_t$  and  $\hat{V}_{MA(1)}$  based on 5000 days results for different data generating processes. The price is contaminated with white noise.

Table 7: RMSE of  $RV_t$ ,  $RK_t^N$ ,  $\hat{V}_t$ ,  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  (Dependent Microstructure Error Case)

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	0.21825	0.39266	0.41585	0.58520
	$RK_t^N$	0.23575	0.44343	0.45080	0.89975
	$\hat{V}_t$	<b>0.21505</b>	<b>0.38582</b>	<b>0.40767</b>	<b>0.54581</b>
	$\hat{V}_{t,MA(1)}$	0.22493	0.40316	0.42100	0.83691
	$\hat{V}_{t,MA(2)}$	0.29260	0.54051	0.57714	1.18875
1-minute	$RV_t$	0.84121	1.48399	1.60189	1.6954
	$RK_t^N$	0.14158	0.25780	0.26987	0.52030
	$\hat{V}_t$	0.84318	1.48663	1.60261	1.67740
	$\hat{V}_{t,MA(1)}$	<b>0.11558</b>	<b>0.20443</b>	<b>0.21297</b>	<b>0.50769</b>
	$\hat{V}_{t,MA(2)}$	0.13732	0.24891	0.26161	0.62325
30-second	$RV_t$	1.66229	2.95397	3.19560	3.37090
	$RK_t^N$	0.11918	0.21559	0.22306	0.42729
	$\hat{V}_t$	1.66480	2.95765	3.19689	3.36058
	$\hat{V}_{t,MA(1)}$	<b>0.08889</b>	<b>0.15848</b>	<b>0.16931</b>	<b>0.34572</b>
	$\hat{V}_{t,MA(2)}$	0.10531	0.18916	0.19313	0.39269
10-second	$RV_t$	4.40694	7.81961	8.49852	7.85934
	$RK_t^N$	0.09850	0.18004	0.18376	0.34594
	$\hat{V}_t$	4.41003	7.82481	8.49935	7.85507
	$\hat{V}_{t,MA(1)}$	0.16456	0.30833	0.30465	0.89045
	$\hat{V}_{t,MA(2)}$	<b>0.06940</b>	<b>0.12804</b>	<b>0.13592</b>	<b>0.25182</b>

This table reports the root mean squared error (RMSE) of estimating 5000 daily ex-post variances using  $RV_t$ ,  $RK_t^N$  and Bayesian nonparametric estimators  $\hat{V}_t$ ,  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  based on returns at different frequencies and simulated from 4 DGPs. The observed prices contains microstructure noise that is dependent with returns.

Table 8: Bias of  $RV_t$ ,  $RK_t^N$ ,  $\hat{V}_t$ ,  $\hat{V}_{t,MA(1)}$  and  $\hat{V}_{t,MA(2)}$  (Dependent Microstructure Error Case)

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	0.16032	0.28455	0.30733	0.17262
	$RK_t^N$	<b>0.01349</b>	0.02985	0.03232	0.00733
	$\hat{V}_t$	0.16005	0.28359	0.30534	0.16275
	$\hat{V}_{t,MA(1)}$	0.01471	<b>0.02665</b>	<b>0.02819</b>	<b>-0.00104</b>
	$\hat{V}_{t,MA(2)}$	0.05581	0.10305	0.11604	0.03956
1-minute	$RV_t$	0.81057	1.42504	1.54563	0.87166
	$RK_t^N$	0.02421	0.04351	0.04360	0.01839
	$\hat{V}_t$	0.81269	1.42805	1.54689	0.86954
	$\hat{V}_{t,MA(1)}$	<b>0.00822</b>	<b>0.01401</b>	<b>0.01359</b>	-0.01044
	$\hat{V}_{t,MA(2)}$	0.01694	0.03179	0.02977	<b>-0.00588</b>
30-second	$RV_t$	1.61481	2.85837	3.10192	1.72912
	$RK_t^N$	0.02791	0.04940	0.05114	0.02369
	$\hat{V}_t$	1.61731	2.86219	3.10359	1.72853
	$\hat{V}_{t,MA(1)}$	<b>0.00721</b>	<b>0.01253</b>	<b>0.00856</b>	-0.01302
	$\hat{V}_{t,MA(2)}$	0.01074	0.01972	0.01796	<b>-0.01155</b>
10-second	$RV_t$	4.32800	7.65381	8.34221	4.67328
	$RK_t^N$	0.04034	0.07209	0.07321	0.04327
	$\hat{V}_t$	4.33106	7.65902	8.34351	4.67462
	$\hat{V}_{t,MA(1)}$	0.11026	0.20188	0.20173	0.13648
	$\hat{V}_{t,MA(2)}$	<b>0.00634</b>	<b>0.01300</b>	<b>0.00850</b>	<b>-0.01896</b>

This table reports the bias estimates from 5000 daily ex-post variances using  $RV$ ,  $RK^N$  and Bayesian nonparametric estimators  $\hat{V}$ ,  $\hat{V}_{MA(1)}$  and  $\hat{V}_{MA(2)}$  based on returns at different frequencies and simulated from 4 DGPs. The observed prices contains microstructure noise that is dependent with returns.

Table 9: Coverage Probability (Dependent Microstructure Error Case)

Data Freq.	Interval Estimator	GARCH	SV1F	SV1FJ	SV2F
5-minute	$RV_t$	76.22%	74.00%	73.12%	21.14%
	$RK_t^N$ - Infeasible	87.26%	87.62%	87.64%	76.72%
	$RK_t^N$ - Feasible	91.16%	91.34%	92.02%	96.42%
	$\hat{V}_t$	66.00%	63.62%	62.44%	16.74%
	$\hat{V}_{t,MA(1)}$	93.96%	94.26%	94.28%	89.84%
	$\hat{V}_{t,MA(2)}$	94.36%	94.60%	94.22%	90.06%
1-minute	$RV_t$	0.00%	0.00%	0.10%	0.06%
	$RK_t^N$ - Infeasible	90.02%	90.40%	89.98%	71.70%
	$RK_t^N$ - Feasible	99.80%	99.80%	99.70%	99.46%
	$\hat{V}_t$	0.00%	0.00%	0.04%	0.04%
	$\hat{V}_{t,MA(1)}$	94.64%	94.92%	94.72%	87.08%
	$\hat{V}_{t,MA(2)}$	94.58%	94.92%	94.30%	86.92%
30-second	$RV_t$	0.00%	0.00%	0.00%	0.00%
	$RK_t^N$ - Infeasible	91.50%	91.72%	91.26%	70.94%
	$RK_t^N$ - Feasible	100.00%	100.00%	100.00%	99.96%
	$\hat{V}$	0.00%	0.00%	0.00%	0.00%
	$\hat{V}_{MA(1)}$	95.00%	95.24%	94.76%	85.18%
	$\hat{V}_{MA(2)}$	94.96%	94.66%	94.78%	85.80%
10-second	$RV_t$	0.00%	0.00%	0.00%	0.00%
	$RK_t^N$ - Infeasible	91.90%	92.44%	92.30%	69.72%
	$RK_t^N$ - Feasible	100.00%	100.00%	100.00%	100.00%
	$\hat{V}_t$	0.00%	0.00%	0.00%	0.00%
	$\hat{V}_{t,MA(1)}$	64.70%	65.00%	68.00%	78.74%
	$\hat{V}_{t,MA(2)}$	94.48%	95.20%	95.14%	82.06%

This table reports the coverage probabilities of 95% confidence intervals of  $RV$ ,  $RK^N$  and 0.95 density intervals of Bayesian nonparametric estimators  $\hat{V}$ ,  $\hat{V}_{MA(1)}$  and  $\hat{V}_{MA(2)}$  based on 5000 days results. The observed prices contains microstructure noise that is dependent with returns.



Table 10: RMSE of  $RV_t$ ,  $RK_t^F$ ,  $\hat{V}_t$  and  $\hat{V}_{t,MA(1)}$  in Tick Time

Data Freq.	Estimator	GARCH	SV1F	SV1FJ	SV2F
<i>Panel A: No Microstructure Noise</i>					
60-tick	$RV_t$	0.11566	0.21887	0.21663	0.49652
	$\hat{V}_t$	<b>0.111175</b>	<b>0.211175</b>	<b>0.21081</b>	<b>0.47565</b>
12-tick	$RV_t$	0.05237	0.09744	<b>0.10194</b>	0.27255
	$\hat{V}_t$	<b>0.05184</b>	<b>0.09661</b>	0.10474	<b>0.26895</b>
6-tick	$RV_t$	0.03866	0.07047	<b>0.07393</b>	0.17821
	$\hat{V}_t$	<b>0.03842</b>	<b>0.07013</b>	0.07787	<b>0.17642</b>
<i>Panel B: Independent Microstructure Noise</i>					
60-tick	$RK_t^F$	0.23958	0.43907	0.44927	0.87832
	$\hat{V}_{t,MA(1)}$	<b>0.20988</b>	<b>0.39083</b>	<b>0.39199</b>	<b>0.69395</b>
12-tick	$RK_t^F$	0.11549	0.20695	0.20794	0.64542
	$\hat{V}_{t,MA(1)}$	<b>0.10545</b>	<b>0.18643</b>	<b>0.18875</b>	<b>0.40857</b>
6-tick	$RK_t^F$	0.08666	0.15397	0.15639	0.31230
	$\hat{V}_{t,MA(1)}$	<b>0.08080</b>	<b>0.13965</b>	<b>0.14555</b>	<b>0.25059</b>

This table reports the RMSE from 5000 daily ex-post variances using  $RV$  and  $\hat{V}$  in no microstructure noise case,  $RK^F$  and  $\hat{V}_{MA(1)}$  in independent noise case, based on tick-time sampled returns.

Table 11: RMSE of  $\hat{V}_t$  and  $\hat{V}_{t,MA(1)}$  in Calendar Time and Tick Time

Estimator	GARCH	SV1F	SV1FJ	SV2F
<i>Panel A: No Microstructure Noise</i>				
5-minute $\hat{V}_t$	<b>0.11125</b>	<b>0.20625</b>	<b>0.20928</b>	<b>0.45238</b>
60-tick $\hat{V}_t$	0.11175	0.21175	0.21081	0.47565
1-minute $\hat{V}_t$	<b>0.05069</b>	<b>0.09195</b>	<b>0.10075</b>	<b>0.18466</b>
12-tick $\hat{V}_t$	0.05184	0.09661	0.10474	0.26895
30-second $\hat{V}_t$	<b>0.03667</b>	<b>0.06597</b>	<b>0.07485</b>	<b>0.14292</b>
6-tick $\hat{V}_t$	0.03842	0.07013	0.07787	0.17642
<i>Panel B: Independent Microstructure Noise</i>				
5-minute $\hat{V}_{t,MA(1)}$	0.21020	<b>0.38103</b>	0.40211	0.73568
60-tick $\hat{V}_{t,MA(1)}$	<b>0.20988</b>	0.39083	<b>0.39199</b>	<b>0.69395</b>
1-minute $\hat{V}_{t,MA(1)}$	<b>0.10304</b>	0.19050	0.19861	<b>0.40527</b>
12-tick $\hat{V}_{t,MA(1)}$	0.10545	<b>0.18643</b>	<b>0.18875</b>	0.40857
30-second $\hat{V}_{t,MA(1)}$	<b>0.07847</b>	0.14306	0.15133	0.28697
6-tick $\hat{V}_{t,MA(1)}$	0.08080	<b>0.13965</b>	<b>0.14555</b>	<b>0.25059</b>

This table reports the root mean squared error (RMSE) of estimating 5000 daily ex-post variances using Bayesian nonparametric volatility estimator  $\hat{V}$  and  $\hat{V}_{MA(1)}$  in calendar time and tick time.

Table 12: Summary Statistics: IBM

Frequency	Data	Mean	Median	Var	Skew.	Kurt.	Min	Max
Daily	$r_t$	0.0673	0.0656	1.6046	0.2069	8.3059	-6.4095	12.2777
	$r_t^2$	1.6091	0.4352	18.9654	13.9087	387.4812	0.0000	150.7429
5-minute	$RV_t$	1.8353	0.9458	11.9867	9.5887	148.2622	0.1032	76.2901
	$RV_{t,M=26}^{block}$	1.8403	0.9506	12.1820	9.7116	151.8262	0.1003	77.2906
	$RK_t^F$	1.6613	0.8447	9.3647	8.5539	124.8480	0.0375	71.9626
	$RK_t^N$	1.6670	0.8476	8.8872	8.0467	109.1098	0.0556	66.3995
	$\hat{V}_t$	1.7805	0.9286	10.3994	8.3839	116.7700	0.1068	70.2477
	$\hat{V}_{t,MA(1)}$	1.6656	0.8424	9.2105	7.2318	77.7981	0.0275	52.3102
	$\hat{V}_{t,MA(2)}$	1.6969	0.8467	10.0917	8.4800	118.7351	0.0137	72.2059
1-minute	$RV_t$	2.0004	1.0468	13.5019	10.5704	202.6835	0.1535	103.8773
	$RV_{t,M=78}^{block}$	2.0045	1.0478	13.6307	10.6737	206.9232	0.1551	105.0582
	$RK_t^F$	1.7952	0.9163	10.8043	8.3092	113.5727	0.1006	73.8576
	$RK_t^N$	1.7425	0.8973	9.6499	7.7187	94.7830	0.0897	60.2024
	$\hat{V}_t$	1.9649	1.0322	12.9922	10.6422	206.7584	0.1517	102.6389
	$\hat{V}_{t,MA(1)}$	1.8417	0.9211	11.3720	7.6213	87.5668	0.1156	64.1797
	$\hat{V}_{t,MA(2)}$	1.7894	0.8979	10.9147	8.5039	121.6750	0.1040	74.7890
	$\hat{V}_{t,MA(3)}$	1.7393	0.8824	9.6571	7.8283	101.5650	0.0986	61.4764
$\hat{V}_{t,MA(4)}$	1.7105	0.8704	9.1269	7.3825	84.7413	0.0964	57.2552	

This table reports the summary statistics of ex-post variance estimators based on 5-minute and 1-minute returns, along with the summary statistics of daily return and daily squared return. The number of daily observation is 3764.

Table 13: HAR and HARQ Model Regression Result Based on IBM Ex-post Variance Estimators

Data Freq.	Parameter	HAR		HARQ	
		$RK_t^F$	$\hat{V}_{t,MA(1)}$	$RK_t^F$	$\hat{V}_{t,MA(1)}$
5-minute	$\beta_0$	0.1322 (0.0374)	0.1224 (0.0375)	0.1015 (0.0382)	-0.0142 (0.0393)
	$\beta_1$	0.1926 (0.0196)	0.2506 (0.0196)	0.2341 (0.0224)	0.4629 (0.0283)
	$\beta_2$	0.5649 (0.0332)	0.4802 (0.0329)	0.5664 (0.0331)	0.4298 (0.0328)
	$\beta_3$	0.1598 (0.0281)	0.1927 (0.0282)	0.1422 (0.0289)	0.1482 (0.0281)
	$\beta_{1Q}$	-	-	-0.0012 (0.0003)	-0.0202 (0.0020)
	R-squared	57.74%	59.55%	57.90%	60.66%
Data Freq.	Parameter	HAR		HARQ	
		$RK_t^N$	$\hat{V}_{t,MA(4)}$	$RK_t^N$	$\hat{V}_{t,MA(4)}$
1-minute	$\beta_0$	0.1246 (0.0365)	0.1308 (0.0376)	0.0065 (0.0367)	-0.0402 (0.0388)
	$\beta_1$	0.2493 (0.0195)	0.2455 (0.0196)	0.4464 (0.0242)	0.5294 (0.0284)
	$\beta_2$	0.5435 (0.0318)	0.5198 (0.0321)	0.5033 (0.0312)	0.4521 (0.0317)
	$\beta_3$	0.1331 (0.0265)	0.1558 (0.0271)	0.0708 (0.0263)	0.0821 (0.0270)
	$\beta_{1Q}$	-	-	-0.0031 (0.0002)	-0.0334 (0.0025)
R-squared	62.71%	60.34%	64.39%	62.19%	

<sup>1</sup> This table reports OLS regression results for the HAR and HARQ model. The results in top panel are based on  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  calculated using 5-minute returns and the bottom panel shows the results of 1-minute  $RK_t^N$  and  $\hat{V}_{t,MA(4)}$ . The values in brackets are standard error of coefficients.

<sup>2</sup> Sample period: 2001/01/03 - 2016/02/16, 3764 observations.

Table 14: Out-of-Sample Forecasts of IBM Volatility

Panel A: 5-minute Return			
Dependent Variable	Regressors	HAR	HARQ
5-minute $RK_t^F$	$RK_t^F$	1.84113	1.84444
	$\hat{V}_{t,MA(1)}$	<b>1.84042</b>	<b>1.81152</b>
5-minute $\hat{V}_{t,MA(1)}$	$RK_t^F$	1.86130	1.86642
	$\hat{V}_{t,MA(1)}$	<b>1.85546</b>	<b>1.83054</b>
Panel B: 1-minute Return			
Dependent Variable	Regressors	HAR	HARQ
1-minute $RK_t^N$	$RK_t^N$	1.87539	<b>1.82881</b>
	$\hat{V}_{t,MA(4)}$	<b>1.87215</b>	1.82548
1-minute $\hat{V}_{t,MA(4)}$	$RK_t^N$	1.94106	1.88974
	$\hat{V}_{t,MA(4)}$	<b>1.93202</b>	<b>1.87276</b>

<sup>1</sup> This table reports the root mean squared forecast error (RMSFE) of forecasting next period ex-post variance using both classical and Bayesian nonparametric variance estimator. Both HAR and HARQ model are considered. The forecasting target is the dependent variable one period out-of-sample.

<sup>2</sup> On each day, the model parameters are re-estimated using all the data up to that day.

<sup>3</sup> Out of sample period: 2005/01/03 - 2016/02/16, 2773 days.

Table 15: Summary Statistics: SPY

Frequency	Data	Mean	Median	Var	Skew.	Kurt.	Min	Max
Daily	$r_t$	0.0188	0.0452	0.4980	-0.6969	6.5418	-4.2837	2.7084
	$r_t^2$	0.4984	0.1634	1.3633	8.8640	118.4977	0.0000	18.3506
5-minute	$RV_t$	0.5287	0.2959	1.5263	16.2802	320.3198	0.0358	25.2066
	$RV_{t,M=26}^{block}$	0.5297	0.2920	1.5736	16.4718	325.8798	0.0356	25.6964
	$RK_t^F$	0.4900	0.2754	0.5812	7.8004	97.4299	0.0094	11.5382
	$RK_t^N$	0.4917	0.2792	0.6305	8.7319	119.7854	0.0168	12.7330
	$\hat{V}_t$	0.5212	0.2901	1.3682	15.7278	304.5983	0.0367	23.5902
	$\hat{V}_{t,MA(1)}$	0.5245	0.2756	1.3761	15.3403	294.2938	0.0160	23.4621
	$\hat{V}_{t,MA(2)}$	0.5162	0.2900	0.8368	10.9368	175.3430	0.0145	16.1662
1-minute	$RV_t$	0.5209	0.3154	0.5902	9.3401	138.3005	0.0521	12.8912
	$RV_{t,M=78}^{block}$	0.5214	0.3174	0.5910	9.3483	138.5217	0.0520	12.9052
	$RK_t^F$	0.5126	0.3036	0.8742	12.6004	219.7084	0.0344	17.4842
	$RK_t^N$	0.5062	0.2999	0.8818	13.0819	232.7427	0.0281	17.8050
	$\hat{V}_t$	0.5170	0.3153	0.5766	9.2625	136.6305	0.0522	12.7063
	$\hat{V}_{t,MA(1)}$	0.5095	0.2984	0.9136	13.1729	235.1810	0.0383	18.1655
	$\hat{V}_{t,MA(2)}$	0.5025	0.2940	0.9899	13.5608	244.4456	0.0344	19.0570
30-second	$RV_t$	0.5034	0.3099	0.4525	7.2251	89.5632	0.0563	10.1295
	$RV_{t,M=130}^{block}$	0.5297	0.2920	1.5736	16.471	325.8798	0.0356	25.6964
	$RK_t^F$	0.5092	0.3109	0.6869	10.9490	177.4869	0.0411	14.7520
	$RK_t^N$	0.5066	0.3094	0.7906	12.2661	211.2367	0.0356	16.4858
	$\hat{V}_t$	0.5008	0.3087	0.4466	7.1778	88.5730	0.0566	10.0356
	$\hat{V}_{t,MA(1)}$	0.5007	0.3017	0.6056	9.9364	152.7216	0.0436	13.3641
	$\hat{V}_{t,MA(2)}$	0.4997	0.2917	0.6780	10.6858	170.9371	0.0394	14.5134
10-second	$RV_t$	0.4984	0.3103	0.4172	7.0322	85.1760	0.0741	9.6097
	$RV_{t,M=260}^{block}$	0.4984	0.3106	0.4170	7.0330	85.2038	0.0741	9.6077
	$RK_t^F$	0.5021	0.3165	0.5179	8.7480	124.3016	0.0533	11.7691
	$RK_t^N$	0.5051	0.3128	0.5845	9.6693	145.9337	0.0447	12.9900
	$\hat{V}_t$	0.4962	0.3099	0.4109	6.9512	83.4540	0.0741	9.4866
	$\hat{V}_{t,MA(1)}$	0.4921	0.3067	0.4220	6.9830	83.8263	0.0470	9.6152
	$\hat{V}_{t,MA(2)}$	0.4943	0.3066	0.4271	6.9312	83.1514	0.0441	9.6627

This table reports the summary statistics of ex-post variance estimators based on 5-minute, 1-minute, 30-second and 10-second SPY returns, along with the summary statistics of daily return and daily squared return. Sample period: 07/02/2014 - 06/28/2016.

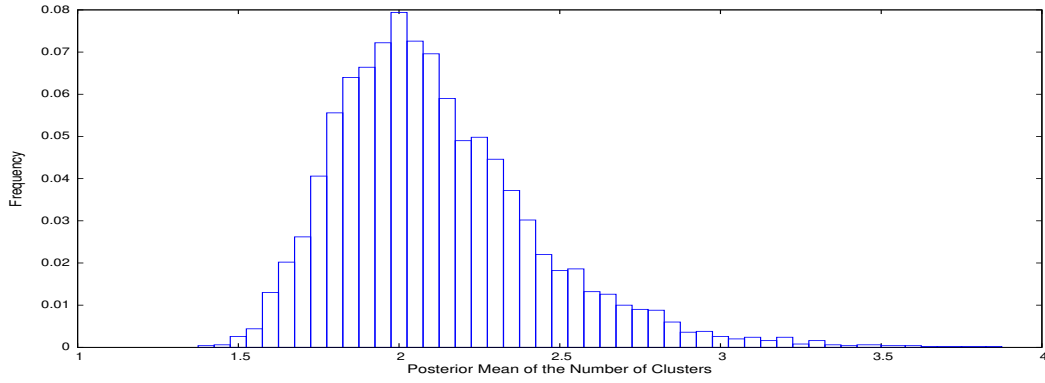


Figure 1: Posterior Mean of the Number of Clusters. Model: DPM. Data: 5-minute return without microstructure noise from SV1F.

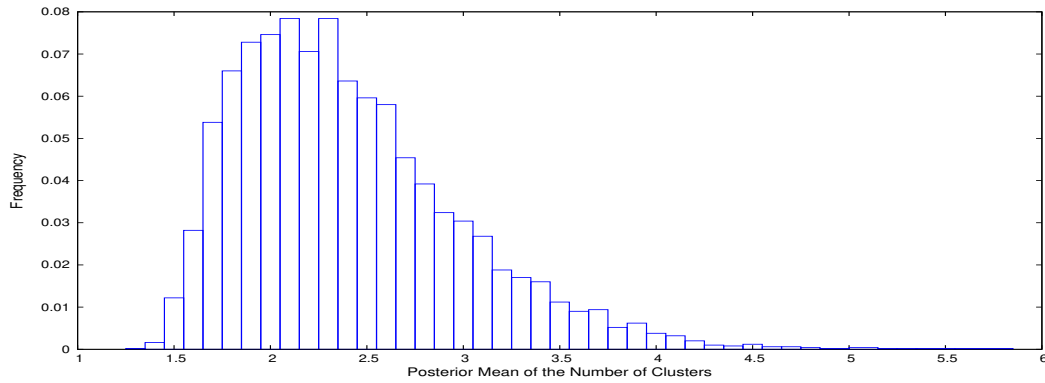


Figure 2: Posterior Mean of the Number of Clusters. Model: DPM-MA(1). Data: 1-minute return with independent noise from SV1FJ

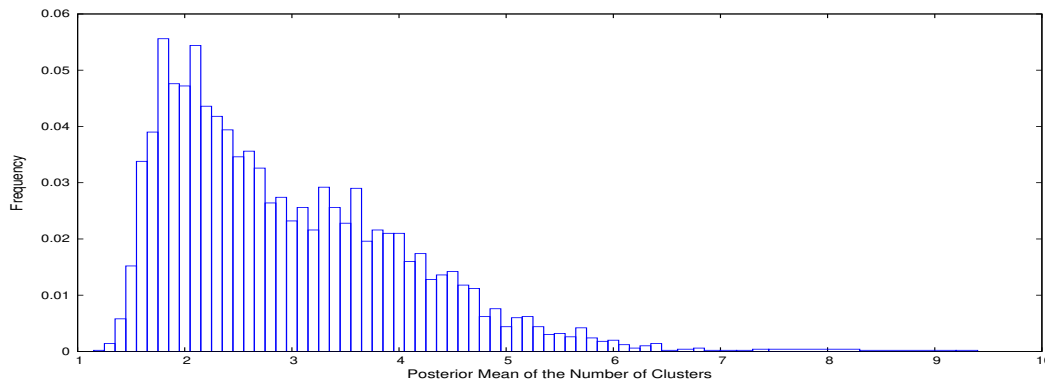


Figure 3: Posterior Mean of the Number of Clusters. Model: DPM-MA(2). Data: 30-second return with dependent noise from SV2F.

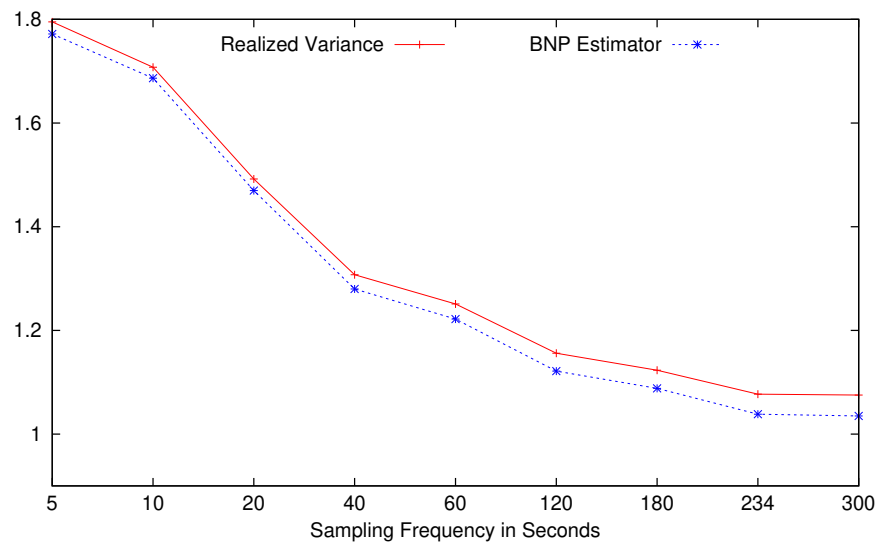


Figure 4: Signature Plot of  $RV_t$  and  $\hat{V}_t$  (IBM data)



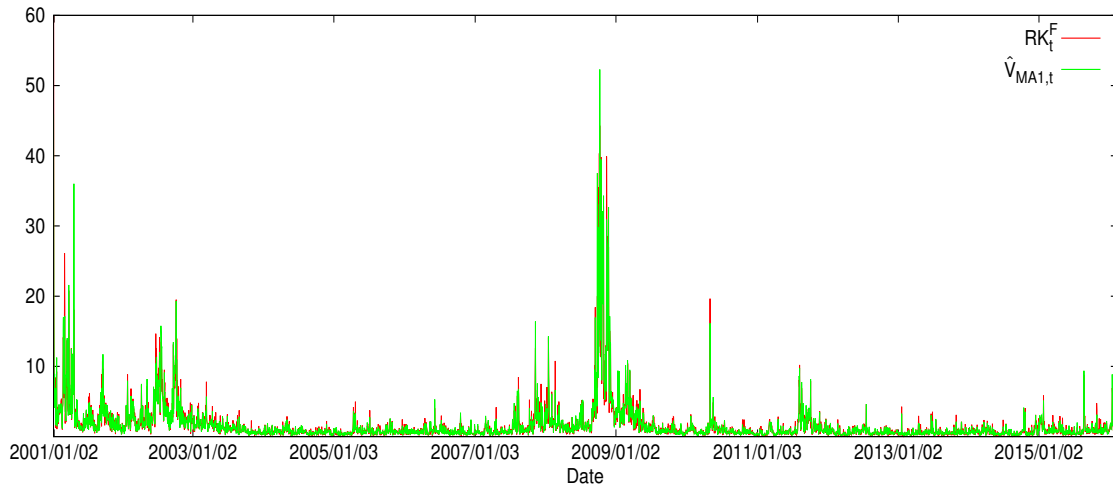


Figure 5:  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  based on 5-minute IBM returns

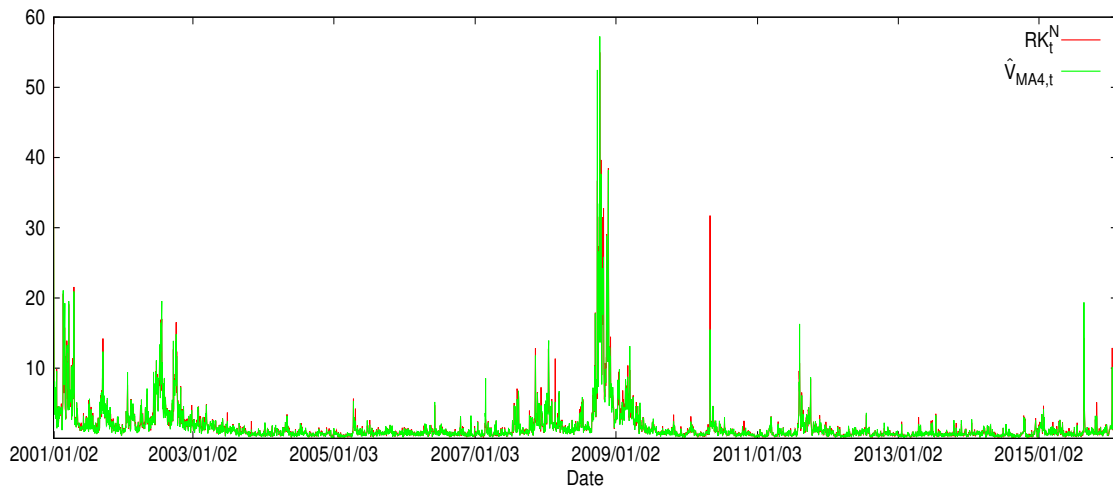


Figure 6:  $RK_t^N$  and  $\hat{V}_{t,MA(4)}$  based on 1-minute IBM returns



Figure 7: High volatility period:  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  calculated using 5 minute IBM returns. Top: variance, below: log-variance

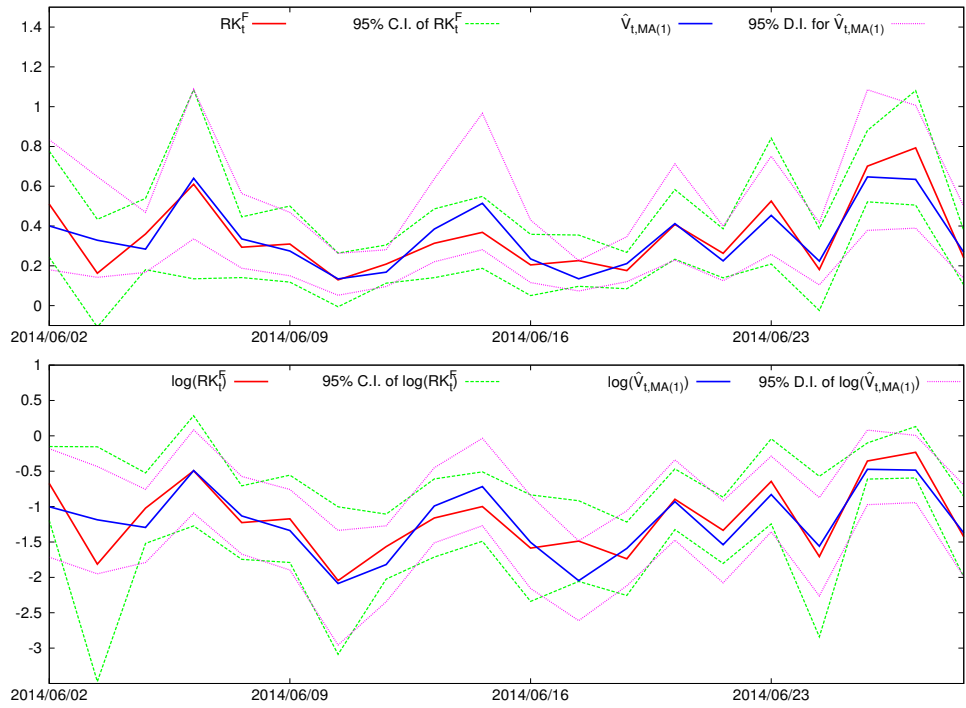


Figure 8: Low volatility period:  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  calculated using 5 minute IBM returns. Top: variance, below: log-variance

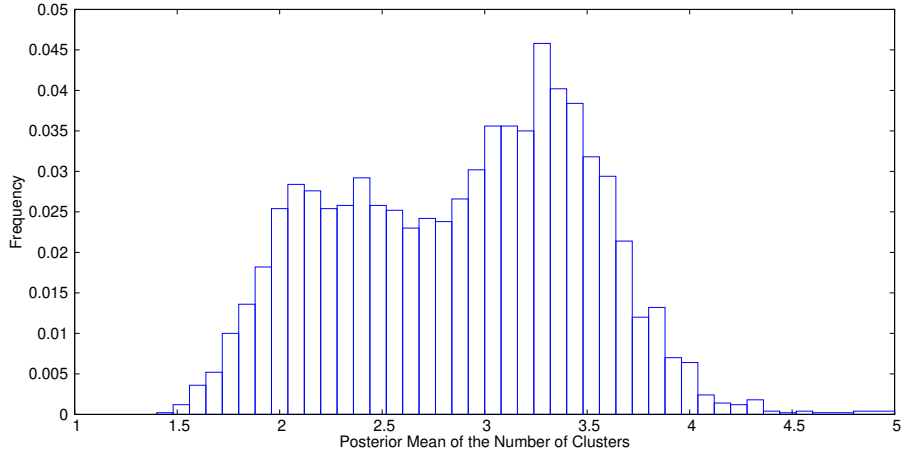


Figure 9: Posterior Mean of the Number of Clusters (Based on 3764 days results from DPM-MA(1) using 5-minute IBM returns).

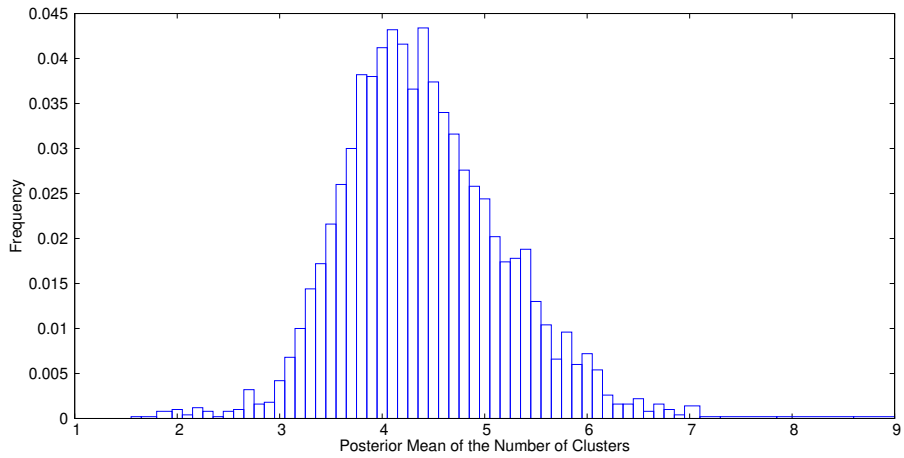


Figure 10: Posterior Mean of the Number of Clusters (Based on 3764 days results from DPM-MA(4) using 1-minute IBM returns).

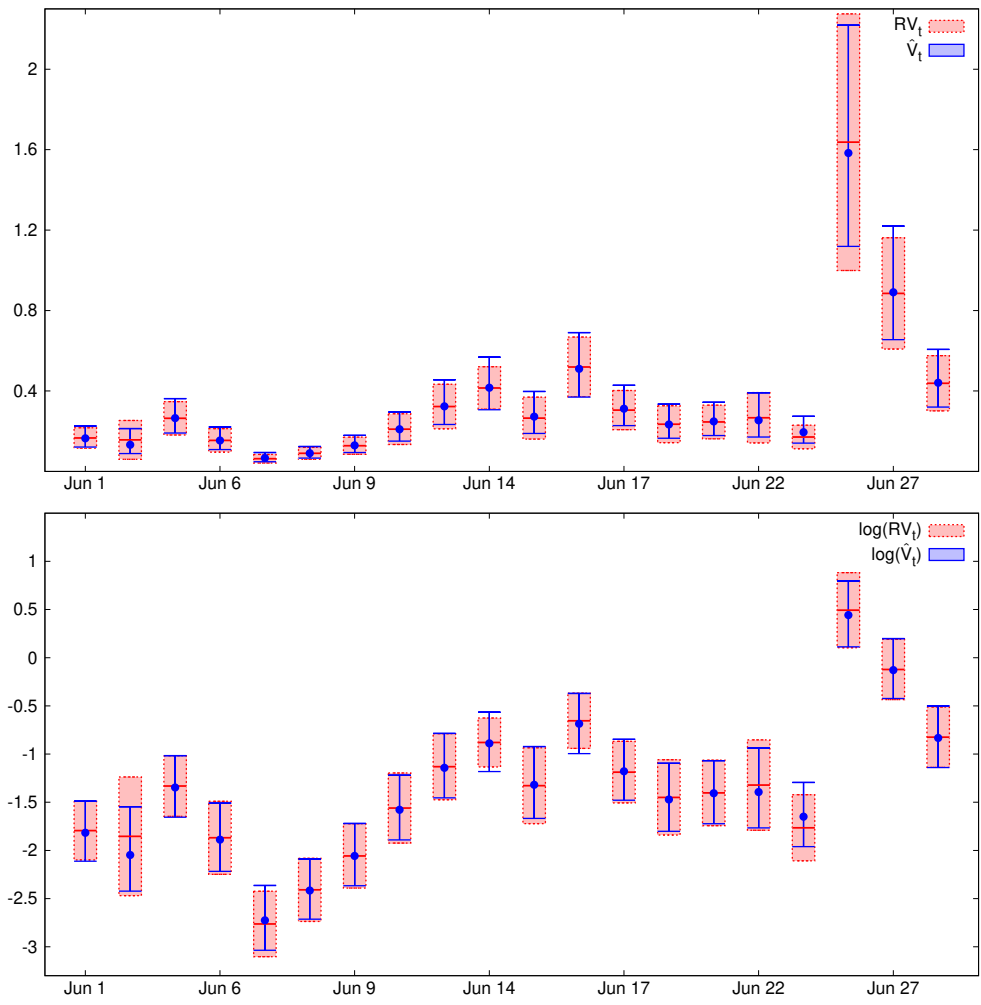


Figure 11:  $RV_t$  and  $\hat{V}_t$  based on 5-minute SPY returns in June 2016. Top: variance, below: log-variance

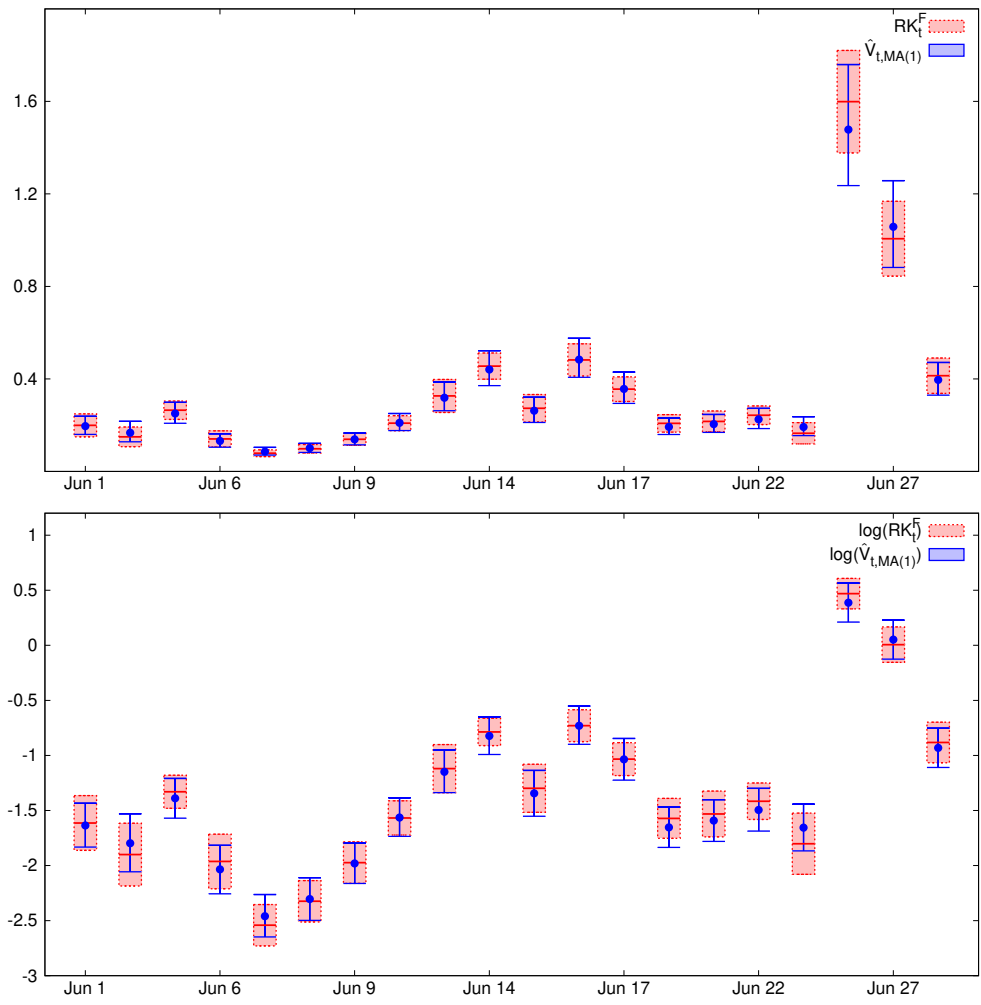


Figure 12:  $RK_t^F$  and  $\hat{V}_{t,MA(1)}$  based on 30-second SPY returns in June 2016. Top: variance, below: log-variance