Identifying Jumps in Financial Assets: a Comparison between Nonparametric Jump Tests [Extended Version] *

March, 2011

Revised: October, 2011

Ana -Maria DUMITRU

Department of Economics and Technology Management, University of Bergamo (Italy) &

Centre for Econometric Analysis, Faculty of Finance, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ (UK).E-mail Ana.Dumitru.1@city.ac.uk

Giovanni URGA

Centre for Econometric Analysis, Faculty of Finance, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ (UK). E-mail: g.urga@city.ac.uk &

Hyman P. Minsky Department of Economic Studies, University of Bergamo (Italy)

Abstract

We perform a comprehensive Monte Carlo comparison between nine procedures available in the literature to detect jumps in financial assets proposed by Barndorff-Nielsen and Shephard (2006), Andersen et al. (2007), Lee and Mykland (2008), Aït-Sahalia and Jacod (2008), Jiang and Oomen (2008), Andersen et al. (2009) (two tests), Corsi et al. (2010) and Podolskij and Ziggel (2010). We evaluate size and power properties of the procedures under alternative sampling frequencies, levels of volatility, persistence in volatility, degree of contamination with microstructure noise, jump size and intensity. The overall best performance is showed by the Lee and Mykland (2008) and Andersen et al. (2007) intraday procedures, provided the price process is not not very volatile. We propose an improvement to these procedures based on critical values obtained from finite sample approximations of the distribution of the test statistics. We show the validity to use reunion and intersection across procedures and across sampling frequencies for potential users of the tests to minimize spurious jump detection. Finally, we report an empirical analysis using real high frequency data on five stocks listed in the New York Stock Exchange.

Keywords: jumps, nonparametric tests, high frequency data, stochastic volatility, Monte Carlo simulations

JEL classification: C01, C14, C15

*A shorter and revised version of this paper is forthcoming in the *Journal of Business and Economic Statistics*.

1 INTRODUCTION

There is a large consensus in the financial literature, theoretical and applied, that modeling return dynamics requires the specification of a stochastic volatility component, which accommodates the persistence in volatility, and of a jump component, which takes care of the unpredictable, large movements in the price process. The identification of the time and the size of jumps has profound implications in risk management, portfolio allocation, derivatives pricing (Aït-Sahalia, 2004). For this task, the use of jump diffusion models proved very difficult, as there are no closed forms of the likelihood function and in addition, the number of parameters to estimate is very high. One solution is to focus on the popular class of affine models (Duffie et al., 2000) which allow for tractable estimation, but impose a quite restrictive set of assumptions. An alternative approach is represented by nonlinear volatility models. However, the estimation procedure, based on simulation methods, such as the Gallant and Tauchen (2002)'s efficient method of moments, is computationally demanding and too much dependent on the choice of an auxiliary model (Chernov et al., 2003; Andersen et al., 2002, see, for instance).

One of the main advances in high frequency econometrics over the last decade was the development of nonparametric procedures to test for the presence of jumps in the path of a price process during a certain time interval or at certain point in time. Such methods are very simple to apply, they just require high frequency transaction prices or mid-quotes. Moreover, they are developed in a model free framework, incorporating different classes of stochastic volatility models. In addition to the seminal contribution of Barndorff-Nielsen and Shephard (2006), in this paper we consider eight other tests proposed by Andersen et al. (2007), Lee and Mykland (2008), Aït-Sahalia and Jacod (2008), Jiang and Oomen (2008), Andersen et al. (2009) (two tests based on the minimum and median realized variance), Corsi et al. (2010) and Podolskij and Ziggel (2010). All tests are based on CLT-type results that require an intraday sampling frequency that tends to infinity. The test statistics are based on robust to jumps measures of variation in the price processes which are estimated by using one of the following types of estimators: realized multi-power variations (Barndorff-Nielsen et al., 2006), threshold estimators (Mancini, 2009), the median or the minimum realized variation (Andersen et al., 2009), the corrected realized threshold multipower variation (Corsi et al., 2010). The Andersen et al. (2007) and Lee and Mykland (2008) tests have the null hypothesis of continuity of the sample path at a certain moment, allowing for the exact identification of the time of a jump. The other procedures have a null of continuity within a certain time period, such as a trading day.

Given such a variety of nonparametric methodologies to identify jumps, one might wonder which

procedure should be preferred, or whether there are data characteristics for which it is recommended to use one test instead of the others. The main objective of this paper is to perform a thorough comparison among the various testing procedures, based on a comprehensive set of Monte Carlo simulations, which embodies important features of financial data. To quantify the size for all tests, our simulations are based on a stochastic volatility model with varying persistence. To evaluate the power property, we consider stochastic volatility models with jumps of different sizes arriving with varying intensity.

Based on the findings of the simulation exercise, we aim to provide a set of guidelines to users of nonparametric tests for jumps. It is important to establish whether the performance of the tests is related to some features of the data, such as different sampling frequencies, different levels of volatility, varying persistence in volatility, varying contamination with microstructure noise, varying jump size and jump intensity. Such characteristics vary between classes of assets, as well as between different time periods. For instance, equity prices are 'jumpier' than bond prices and markets in general have been more volatile and at the same time 'jumpier' during the last three years than before.

We make two additional contributions to the existing literature. First, in the case of the Andersen et al. (2007) and Lee and Mykland (2008) tests, we explore the benefits from using approximate finite sample distributions. We generate critical values based on simulations, in line with White (2000)'s Monte Carlo Reality Check approach. Second, we propose a procedure that combines tests and frequencies to reduce the probability of detecting spurious jumps.

Finally, we apply the tests to high frequency data for five stocks listed in the New York Stock Exchange, namely Procter&Gamble, IBM, JP Morgan, General Electric and Disney, during 2005 and 2009.

To the best of our knowledge, in the literature there are two other papers that deal with similar issues. Theodosiou and Žikeš (2010) perform an extensive Monte Carlo simulation exercise to evaluate the performance of different jump detection procedures, with a special interest in the effect of illiquid data on the behaviour of the various tests. Schwert (2009) instead relies only on real data to conclude that different jump detection procedures pick up different jumps. Our paper is more comprehensive in terms of testing procedures included in our comparison. In addition, while we acknowledge that tests for jumps can lead to very different findings, however we provide a feasible solution to this problem first, by proposing the use of simulated critical values for the Andersen et al. (2007) and Lee and Mykland (2008) tests; second, and most importantly, we show that combining various procedures

greatly improves the performance of the tests in terms of spurious jump detection.

The paper is organized as follows. In Section 2, we review the nine nonparametric tests for jumps available in the literature. Section 3 describes the Monte Carlo setup and reports the main findings of the simulations. Section 4 reports on the extensions to the existing tests based on approximations of the finite sample distributions of the test statistics for the intraday procedures and the benefits from combinations of the existing tests. Section 5 reports an empirical exercise using stock data. Finally, Section 6 concludes and offers some guidelines to potential users.

2 JUMP TESTS

In this section, we describe the available jump detection procedures. First, let us briefly illustrate the theoretical framework in which all tests have been developed.

The logarithmic price process, p_t, is usually assumed to be a jump-diffusion process of the form:

$$\mathrm{d}p_t = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}W_t + \mathrm{d}J_t \tag{1}$$

where μ_t represents the drift, σ_t the diffusion parameter, and W_t a Brownian motion at time t. J_t is the jump process at time t, defined as $J_t = \sum_{j=1}^{N_t} c_{t_j}$. c_{t_j} represents the size of the jump at time t_j and N_t is a counting process, representing the number of jumps up to time t.

The quadratic variation (QV) of the price process up to a certain point in time t (usually a trading day) can be defined as follow:

$$[p]_t = \int_0^t \sigma_s^2 \mathrm{d}s + \sum_{j=1}^{N_t} c_{t_j}^2, \tag{2}$$

where $\int_0^t \sigma_s^2 \mathrm{d}s$ is the integrated variance or volatility (IV). Thus, $[p]_t$ is made up of a part coming from the diffusion component and another one caused by the jump component. The two components have a different nature and should be separately analyzed and modelled. The integrated volatility is characterized by persistence, whereas jumps, apart from a possible drift, have an unpredictable nature.

The recent literature in the field of high frequency econometrics has developed several estimators for both the quadratic variance and the integrated volatility of a price process such as the one derived in (1). Most of these estimators are based on equally spaced data. Thus, the interval $[0, t]$ is split

into n equal subintervals of length δ. The j-th intraday return r_j on day t is defined as follows:

$$r_j = p_{t-1+j\delta} - p_{t-1+(j-1)\delta}. \tag{3}$$

$[p]_t$ can be estimated by the realized variance (RV_t), defined as (Andersen and Bollerslev, 1998):

$$RV_t = \sum_{j=1}^{n} r_j^2 \xrightarrow{\delta \to 0} [p]_t, \tag{4}$$

where $\xrightarrow{\delta \to 0}$ stands for convergence in probability when $\delta \to 0$.

To measure the IV one can use a wide range of estimators, such as multipower variations, threshold estimators, medium and minimum realized variance. All these quantities are robust to jumps in the limit. Most of the jump detection procedures are based on the comparison between RV_t, which captures the variation of the process generated by both the diffusion and the jump parts, and a robust to jumps estimator.

It is important to note that none of these procedures can test for the absence or presence of jumps in the model or data generating process. They merely supply us with information on whether within a certain time interval or at a certain moment, the realization of the process is continuous or not. Andersen et al. (2007) and Lee and Mykland (2008) assume the null of continuity of the sample path at time t_j. For all the other procedures, the null is of continuity of the sample path during a certain period, such as a trading day. The alternative hypothesis implies discontinuity of the sample path, that is the occurrence of at least one jump.

Apart from the procedures proposed by Aït-Sahalia and Jacod (2008) and Podolskij and Ziggel (2010), all other procedures work only when a finite number of jumps occur within a certain time interval. This is due to the fact that in most cases, the construction of the test statistics is based on realized multi-power variation estimators, which are robust only to a finite number of jumps. For this reason, in the simulation set-up, we only consider processes with a finite number of jumps (compound Poisson) and compare tests under this scenario.

In the light that the Andersen et al. (2007) and Lee and Mykland (2008) tests differ only in terms of the choice of the critical values, for a large part of our simulation exercise, we do not distinguish between the two of them (see Section 2.2 and the Remarks in Section 3.1).

We turn now to the presentation of the procedures.

2.1 Barndorff-Nielsen and Shephard (2006) test (BNS henceforth)

Barndorff-Nielsen and Shephard (2006) base their procedure on the possibility to build a consistent estimator for the integrated variance of a process. The test draws from previous research (Barndorff-Nielsen and Shephard, 2004), where authors show that the realized bipower variation (BV_t) consistently estimates the integrated variance in the presence of rare jumps:

$$BV_t = \underset{\delta \downarrow 0}{\text{plim}} \sum_{j=2}^{n} |r_j||r_{j-1}| \tag{5}$$

Barndorff-Nielsen et al. (2006) generalize the BV_t to realized multipower variations, computed as sums of products of adjacent absolute returns raised to certain powers. These quantities can be generally used to estimate $\int_0^t \sigma_s^m \mathrm{d}s$, $m > 0$ in the presence of jumps.

One can infer whether jumps occur during a time interval (usually a trading day) by comparing the realized volatility with the realized bipower variation. Following simulation studies reported by the authors and also by Huang and Tauchen (2005), in this paper we use the ratio test defined as:

$$\frac{1 - \frac{BV_t}{RV_t}}{\sqrt{(\mu_1^{-4} + 2\mu_1^{-2} - 5)\delta \max\left(1, \frac{TQ_t}{BV_t^2}\right)}} \xrightarrow{L} \mathcal{N}(0,1) \tag{6}$$

where $\mu_1 = \sqrt{2/\pi}$ and $\xrightarrow{L}$ stands for convergence in law. TQ_t represents the realized tripower quarticity that consistently estimates the integrated quarticity, i.e. $\int_0^t \sigma_u^4 \, \mathrm{d}u$, and is defined as follows:

$$TQ_t = n\mu_{4/3}^{-3} \left(\frac{n}{n-2}\right) \sum_{j=3}^{n} |r_{j-2}|^{4/3} |r_{j-1}|^{4/3} |r_j|^{4/3} \tag{7}$$

where $\mu_{4/3} = E(|U|)^{4/3}$, with U being a standard normal variable.

2.2 Andersen et al. (2007) and Lee and Mykland (2008) tests (ABD and LM henceforth)

Lee and Mykland (2008) and Andersen et al. (2007) develop tests for jumps based on the standardization of intraday returns by robust to jumps volatility estimators. Both tests are constructed under the null that there is no jump in the realization of the process at a certain time, t_j. This enables users to identify the exact time of a jump, as well as the number of jumps within a trading day. We call these two procedures "intraday" tests, as they can detect jumps that occur any time

during a trading day, whereas the other tests in the literature can only check for the discontinuity of the sample path at a daily level.

The first step in applying both ABD and LM procedures is to compute a local (spot) volatility estimate that is robust to jumps and then standardize the intraday returns with this estimate. Given the intraday return at time t_j, i.e. r_j, and the local volatility estimate, $\hat{V}_j$, authors define the following statistic:

$$z_j = \frac{|r_j|}{\sqrt{\hat{V}_j}} \tag{8}$$

Both papers propose computing $\hat{V}_j$ as the properly scaled realized bipower variation over a window around or before t_j:

$$\hat{V}_j = \frac{BV_{t_j}}{K-2}, \tag{9}$$

where K is the window size on which BV_{t_j} is calculated.

As z_j is proved to be asymptotically normal, one can attempt to identify jumps by comparing it to a normal threshold, as proposed by Andersen et al. (2007). As the test is applied at every intraday time, t_j, in order to deal with the false discovery rate issue which may arise in the context of multiple testing, the authors propose using the Šidák approach. Once a nominal daily size, α, is fixed, the corresponding size for each intraday test is defined as $\beta = 1-(1-\alpha)^{\delta}$. If $z_j > \Phi_{1-\beta/2}$, we reject the null of continuity of the sample path.

Lee and Mykland (2008) use a slightly different approach. The usual 95% and 99% quantiles from the normal distribution prove too permissive, leading to an over-rejection of the null. To overcome this limitation, the authors propose using critical values from the limit distribution of the maximum of the test statistics. They show that this maximum converges, for $\delta \to 0$, to a Gumbel variable:

$$\frac{\max(z_j) - C_n}{S_n} \xrightarrow{L} \xi, \qquad \mathbf{P}(\xi) = \exp(-e^{-x}) \tag{10}$$

where $C_n = \frac{(2\log n)^{1/2}}{\mu_1} - \frac{\log\pi + \log(\log n)}{2\mu_1(2\log n)^{1/2}}$ and $S_n = \frac{1}{\mu_1(2\log n)^{1/2}}$.

The test can be conducted by comparing z_j, standardized as $\max(z_j)$ in (10), to the critical value from the Gumbel distribution.

It is worth noting the following regarding the implementation of the two tests. Andersen et al. (2007) provide no suggestions concerning the sample size on which to estimate the local volatility. Lee and Mykland (2008) instead propose computing $\hat{\sigma}_j$ on a window size of K observations that precede time t_j. They show that K depends on the choice of the sampling frequency and suggest to

take $K = \sqrt{252 * n}$, where n is the daily number of observations, whereas 252 is the number of days in the (financial) year.

The ABD test requires very low nominal sizes (10^{-5}), whereas for all other procedure, we use a 5% significance level. In order to assure comparability with the other procedures, we do not distinguish between the two procedures and use the critical values of Lee and Mykland (2008). Thus, we report the results under the acronym 'ABD-LM'.

Whenever we make comparisons with the other tests which are applied on time intervals equal to one trading day, we compute the 'ABD-LM' test statistics for every moment t_j within a trading day and then pick up the maximum statistic as the final test for that day.

2.3 The Aït-Sahalia and Jacod (2008) test (AJ henceforth)

Another procedure that enables the identification of discontinuities in prices is the one developed by Aït-Sahalia and Jacod (2008). Consider the following realized power variation:

$$B(m, \delta)_t = \sum_{j=1}^{[t/\delta]} |r_j|^m, \tag{11}$$

with the scalar $m > 0$. Aït-Sahalia and Jacod (2008) notice that when $m > 2$ and jumps are present, $B(m, \delta)_t$ is invariant to sampling scale modifications. This is no longer valid for continuous processes. Based on this observation, authors develop a family of test statistics that compare realized power variations computed on data sampled at two different scales, δ and $k\delta$, $k \in \mathbb{N}$. Define $\widehat{S(m, k, \delta)_t}$ as:

$$\widehat{S(m, k, \delta)_t} = \frac{\widehat{B(m, k\delta)_t}}{\widehat{B(m, \delta)_t}} \xrightarrow{\delta \to 0} k^{m/2-1}, \tag{12}$$

where $m > 2$ and $k \geq 2$. The following test statistic is proposed to test for the null of no jumps:

$$\frac{\widehat{S(m, k, \delta)_t} - k^{m/2-1}}{\sqrt{V_{n,t}}} \xrightarrow{L} \mathcal{N}(0, 1), \tag{13}$$

where $V_{n,t}$ is the variance of the test statistic and we refer to the original contribution for details. $V_{n,t}$ can be estimated by using both multipower variations or threshold estimators (Mancini, 2009). In this paper, we employ both methodologies.

2.4 Jiang and Oomen (2008) test (JO henceforth)

Another approach to jump identification is proposed by Jiang and Oomen (2008), with the null of no jumps in the sample path between 0 and t. The test exploits the differences that can occur between arithmetic and logarithmic returns computed as follows:

$$SwV_t(\delta) = 2\sum_{j=1}^{[t/\delta]}(R_j - r_j) \tag{14}$$

where R_j denotes the arithmetic return j-th intraday return, while r_j is the log return. The absence of jumps makes the difference between SwV_t and the realized variance equal to 0:

$$\operatorname*{plim}_{\delta\to 0}(SwV_t - RV_t) = \begin{cases} 0 & \text{no jumps in}[0,t] \\ 2\int_0^t \underline{J}_u \, \mathrm{d}q_u - \int_0^t J_u^2 \, \mathrm{d}q_u & \text{jumps in}[0,t] \end{cases} \tag{15}$$

where $\underline{J}_u = exp(J_u) - J_u - 1$, with J the jump process.

The test statistic is defined as:

$$\frac{nBV_t}{\sqrt{\Omega_{SwV}}}\left(1 - \frac{RV_t}{SwV_t}\right) \xrightarrow{L} \mathcal{N}(0,1). \tag{16}$$

Ω_{SwV} is estimated using a realized multipower variation (Barndorff-Nielsen et al., 2003; Barndorff-Nielsen et al., 2006):

$$\hat{\Omega}_{SwV} = \frac{\mu_6}{9}\frac{n^3\mu_{6/m}^{-m}}{n-m+1}\sum_{i=0}^{n-m}\prod_{k=1}^{m}|r_{i+k}|^{6/m} \tag{17}$$

where a suitable choice for m is either 4 or 6, as suggested by the authors, and $\mu_6 = E(|U|)^6$, $U \sim \mathcal{N}(0,1)$.

2.5 Andersen et al. (2009) tests based on MinRV and MedRV (Min and Med tests henceforth)

Andersen et al. (2009) show that the realized multipower variations are very sensitive to market microstructure noise, especially to zero returns. Authors propose to use instead estimators based on the nearest neighbour truncation. The minimum realized variance ($MinRV_t$) and median realized

variance ($MedRV_t$) are proposed to estimate the integrated volatility in the presence of jumps:

$$\begin{aligned} MinRV_t &= \tfrac{\pi}{\pi-2}\tfrac{n}{n-1}\textstyle\sum_{j=2}^{n}\min(|r_j|,|r_{j-1}|)^2 \\ MedRV_t &= \tfrac{\pi}{6-4\sqrt{3}+\pi}\tfrac{n}{n-2}\textstyle\sum_{j=3}^{n}\mathrm{med}(|r_j|,|r_{j-1}|,|r_{j-2}|)^2. \end{aligned} \tag{18}$$

In line with the BNS procedure, authors construct tests for jumps from the comparison between the above estimators and RV_t:

$$\begin{aligned} &\frac{1-\frac{MinRV_t}{RV_t}}{\sqrt{1.81\ \delta\ \max\left(1,\frac{MinRQ_t}{MinRV_t^2}\right)}} \xrightarrow{L} \mathscr{N}(0,1) \quad \text{and} \\ &\frac{1-\frac{MedRV_t}{RV_t}}{\sqrt{0.96\ \delta\ \max\left(1,\frac{MedRQ_t}{MedRV_t^2}\right)}} \xrightarrow{L} \mathscr{N}(0,1), \end{aligned} \tag{19}$$

where $MinRQ_t = \frac{\pi n}{3\pi-8}\frac{n}{n-1}\sum_{j=2}^{n}\min(|r_j|,|r_{j-1}|)^4$ is the minimum realized quarticity and $MedRQ_t = \frac{3\pi n}{9\pi+72-52\sqrt{3}}\frac{n}{n-2}\sum_{j=3}^{n}\mathrm{med}(|r_j|,|r_{j-1}|,|r_{j-2}|)^4$ the median realized quarticity which estimate the integrated quarticity.

2.6 Corsi et al. (2010) test (CPR henceforth)

Corsi et al. (2010) stress the shortcomings of the realized multipower variations and propose the corrected realized threshold multipower variation. Authors propose the following test statistic:

$$\frac{1-\frac{C-TBV_t}{RV_t}}{\sqrt{\left(\frac{\pi^2}{4}+\pi-5\right)\delta\ \max\left(1,\frac{C-TTriPV_t}{C-TBV_t^2}\right)}} \xrightarrow{L} \mathscr{N}(0,1), \tag{20}$$

where $C-TBV_t$ and $C-TTriPV_t$ represent the corrected realized threshold bipower and tripower variation, respectively, defined as:

$$\begin{aligned} C-TBV_t &= \tfrac{\pi}{2}\textstyle\sum_{j=2}^{n} Z1(r_j,\vartheta_j)Z1(r_{j-1},\vartheta_{j-1}), \\ C-TTriPV_t &= \mu_{4/3}^{-3}\textstyle\sum_{j=3}^{n} Z1(r_j,\vartheta_j)Z1(r_{j-1},\vartheta_{j-1})Z1(r_{j-2},\vartheta_{j-2}) \end{aligned} \tag{21}$$

where $\mu_{4/3} = E(|U|)^{4/3}$, $U \sim \mathscr{N}(0,1)$ and $Z1(r_j,\vartheta_j) = \begin{cases} |r_j|, & r_j^2 < \vartheta_j \\ 1.094\ \vartheta_j^{\frac{1}{2}}, & r_j^2 > \vartheta_j \end{cases}$ is a function of the return at time t_j and a threshold $\vartheta_j = c_\vartheta^2 \cdot \hat{V}_j$. c_ϑ^2 is a scale free constant and $\hat{V}_j$ a local volatility

estimator.

Following authors' recommendation, to compute the threshold, ϑ_j, we take $c_\vartheta = 3$. For the auxiliary local volatility estimate, $\hat{V}_t$, Corsi et al. (2010) propose using a non-parametric filter that removes jumps from data in several iterations. We refer to the original paper (in particular Annex B) for details.

2.7 Podolskij and Ziggel (2010) test (PZ henceforth)

This procedure is based on comparison between a realized power variation and a robust to jumps estimator to detect jumps, as in the case of the BNS, Min, Med and CPR tests. Podolskij and Ziggel (2010)'s choice for the robust to jumps estimator is Mancini (2009)'s threshold estimator. However, since the derivation of a limiting theory for the simple differentiation between the two has proved particularly difficult, authors define the test statistics as a difference between a realized power variation estimator and a threshold estimator perturbed by some external positive i.i.d. random variables, $(\eta_j)_{1\leq j\leq [t/\delta]}$, with $E[\eta_j] = 1$ and finite variation:

$$T(m,\delta)_t = n^{\frac{m-1}{2}} \sum_{j=1}^{[t/\delta]} |r_j|^m (1 - \eta_j \mathbb{I}_{\{|r_j|\leq c\delta^w\}}), \qquad m \geq 2, \tag{22}$$

where $1_{\{|r_j|\leq c*\delta^w\}}$ is an indicator function for absolute returns lower than a threshold fixed to $c * \delta^w$, with $c = 2.3\sqrt{BV_t}$ and $w = .4$.

The test statistic can be defined as follows:

$$\frac{T(m,\delta)_t}{\sqrt{Var[\eta_j] n^{\frac{2m}{2}-1} \sum_{j=1}^{[t/\delta]} |r_j|^{2m} \mathbb{I}_{\{|r_j|\leq c\delta^w\}}}} \xrightarrow{L} \mathcal{N}(0,1), \tag{23}$$

where $Var[\eta_j]$ is the variance of the η_j variables. For the perturbing variables, Podolskij and Ziggel (2010) recommend to sample them from the following distribution:

$$P^\eta = \frac{1}{2}(\varsigma_{1-\tau} + \varsigma_{1+\tau}), \tag{24}$$

where ς is the Dirac measure, and τ is a constant chosen relatively small, e.g. $\tau = 0.1$ or 0.05.

3 MONTE CARLO ANALYSIS

In this section we report and discuss the results of an extensive comparison among the testing procedures presented in the previous section. The exercise is based on a comprehensive set of Monte Carlo simulations, which embody several features of financial data. To quantify the size for all tests, our simulations are based on stochastic volatility models with varying persistence. To evaluate the power property, we consider stochastic volatility models with jumps of different sizes arriving with varying intensity.

3.1 Simulation design

This section provides a description of the Monte Carlo design. Following Huang and Tauchen (2005), we simulated several stochastic volatility processes with leverage effect, with or without jumps and different levels of persistence in volatility, as well as varying jump intensities and jump variances.

The benchmark model for our simulations is a stochastic volatility model with one volatility factor (SV1F). The volatility factor enters the price equation in an exponential form, as suggested in Chernov et al. (2003):

$$\begin{aligned} \mathrm{d}p_t &= \mu \mathrm{d}t + \exp[\beta_0 + \beta_1 \upsilon_t] \mathrm{d}W_{p_t}, \\ \mathrm{d}\upsilon_t &= \alpha_\upsilon \upsilon_t \mathrm{d}t + \mathrm{d}W_{\upsilon_t}, \qquad \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_\upsilon) = \rho \end{aligned} \tag{25}$$

where p_t is the log-price process, the W's are standard Brownian motions, υ_t the volatility factor, μ the drift of the price process, α_υ the drift of the volatility process and ρ the leverage effect. This is the process that we simulate under the null hypothesis of no jumps.

Under the alternative hypothesis of discontinuous sample paths, to the price process in (25) we add a compound Poisson process with jump intensity λ and jump size distributed as $N(0, \sigma^2_{jump})$.

Chernov et al. (2003) show that it is possible to generate similar dynamics with the ones produced by a jump diffusion model by using a two factor stochastic volatility model. A first volatility factor controls for the persistence in the volatility process, while the second factor generates higher tails in a similar manner to a jump process. Moreover, by considering the volatility feedback component for the second factor, the model can sometimes accommodate market conditions even better than jump diffusions, as the volatility of volatility can capture the dynamics of extreme events. Thus, a second

stochastic volatility model (SV2F) is defined as:

$$\begin{aligned} \mathrm{d}p_t &= \mu \mathrm{d}t + s - exp[\beta_0 + \beta_1 \upsilon_{1_t} + \beta_2 \upsilon_{2_t}]\mathrm{d}W_{p_t} \\ \mathrm{d}\upsilon_{1_t} &= \alpha_{\upsilon_1}\upsilon_{1_t}\mathrm{d}t + \mathrm{d}W_{\upsilon_{1_t}} \\ \mathrm{d}\upsilon_{2_t} &= \alpha_{\upsilon_2}\upsilon_{2_t}\mathrm{d}t + [1 + \beta_{\upsilon_2}\upsilon_{2_t}]\mathrm{d}W_{\upsilon_{2_t}} \end{aligned} \tag{26}$$

with $\mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_{\upsilon_1}) = \rho_{p,\upsilon_1}$ and $\mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_{\upsilon_2}) = \rho_{p,\upsilon_2}$.

SV2F can generate extreme returns, without having a jump component. We simulate this model only under the null hypothesis. Our objective is to understand whether the various tests for jumps maintain a reasonable size in extremely volatile periods.

To assess the power of the tests, we augment SV1F with rare compound Poisson jumps, arriving with intensity λ and having normally distributed sizes with mean 0 and standard deviation σ_{jump}.

The values of the parameters of the two stochastic volatility models are the ones in Huang and Tauchen (2005) and are reported, for convenience, in Table 1. Table 1 also reports the values of the jump parameters, λ and σ_{jump}:

SV1F		**SV2F**	
μ	0.030	μ	0.030
β_0	0	β_0	-1.200
β_1	0.125	β_1	0.040
α_υ	{-0.137e^{-2}, -0.100, -1.386}	β_2	1.500
ρ	-0.620	α_{υ_1}	-0.137 e^{-2}
λ	0 - 2	α_{υ_2}	-1.386
σ_{jump}	0 - 2.50 by 0.50	β_{υ_2}	0.250
		ρ_{p,υ_1}	-0.300
		ρ_{p,υ_2}	-0.300

Table 1: Parameter values for the 1 factor stochastic volatility models (SV1F) and for the 2 factor model (SV2F)

In empirical applications it is customary to apply these tests at a daily level, in order to be able to conclude whether jumps occurred during the trading day. Therefore, we evaluate the statistical properties of all jump tests based on data simulated for 10000 trading days, for all models and under both hypotheses of continuity and discontinuity. For the simulation of each path, we use an Euler discretization scheme based on increments of 1 second. We then perform a sampling at 1, 5, 15 and 30 minutes. For comparison purposes, all models with the same number of factors are based on the same Brownian motion(s). For instance, for all the models derived from the SV1F model, we use the same simulated Brownian motions to describe the dynamics of both the price and volatility factor.

Figures 1 and 2 report the simulated daily prices, volatility factors and returns for SV1F with medium mean reversion ($\alpha_v = -0.1$) and SV2F, for 10,000 days, with data sampled every 5 minutes.

We report results using a 5% significance level. The results for alternative significance levels, such as 1%, 0.1% and 0.01%, are in line with the ones at 5%. We report size and size adjusted power.

3.2 Monte Carlo findings

3.2.1 Size and power of the tests for stochastic volatility models

SIZE For SV1F, we consider three alternative values of the mean reversion parameter of the volatility factor. In all cases, the empirical size tends to slightly decrease with the increase in the mean reversion parameter, without affecting the ranking of the tests.

Results are not affected by the values taken by the mean reversion parameter. In this paper we only report the empirical size for the medium mean reversion case (see Table 2). The full set of results is available upon request.

Procedure	1sec	1 min	5 min	15 min	30 min
AJ(threshold)	0.047	0.038	0.031	0.027	0.014
AJ(power var)	0.048	0.046	0.051	0.088	0.150
BNS	0.048	0.054	0.053	0.057	0.063
CPR	0.052	0.055	0.056	0.064	0.075
JO	0.065	0.069	0.086	0.122	0.189
ABD-LM	0.055	0.066	0.074	0.063	0.059
Med	0.051	0.050	0.052	0.053	0.064
Min	0.047	0.046	0.044	0.040	0.035
PZ	0.049	0.065	0.083	0.100	0.121

Table 2: Size of the tests for jumps for the SV1F model with medium mean reversion

In Table 2, if we look at all the sampling frequencies, the biggest size distortion is encountered in the case of the JO test, where, for a 1 second sampling frequency, we have a size equal to 6.5%, which increases even more when the sampling frequency diminishes. A similar pattern can be seen for the PZ procedure, which displays a size close to the nominal one when sampling is performed every second, but then gets rapidly and highly oversized.

The best performance is shown by the Med and BNS tests. Both tests display a size very close to the nominal one at a sampling frequency of second, i.e. 5.1% for the Med and 4.8% for BNS. The size then tends to slowly increase with the decrease in the sampling frequencies. The Min and CPR tests also seem to behave well at higher frequencies, with a size of 5.2% for CPR and 4.7% for the Min test. However, the Min test has a tendency of becoming undersized at lower frequencies, getting to 3.5% at 30 minutes. The CPR procedure becomes oversized with the decrease in the sampling frequency and displays a size equal to 7.5% for 30 minutes data. The intraday ABD - LM procedure

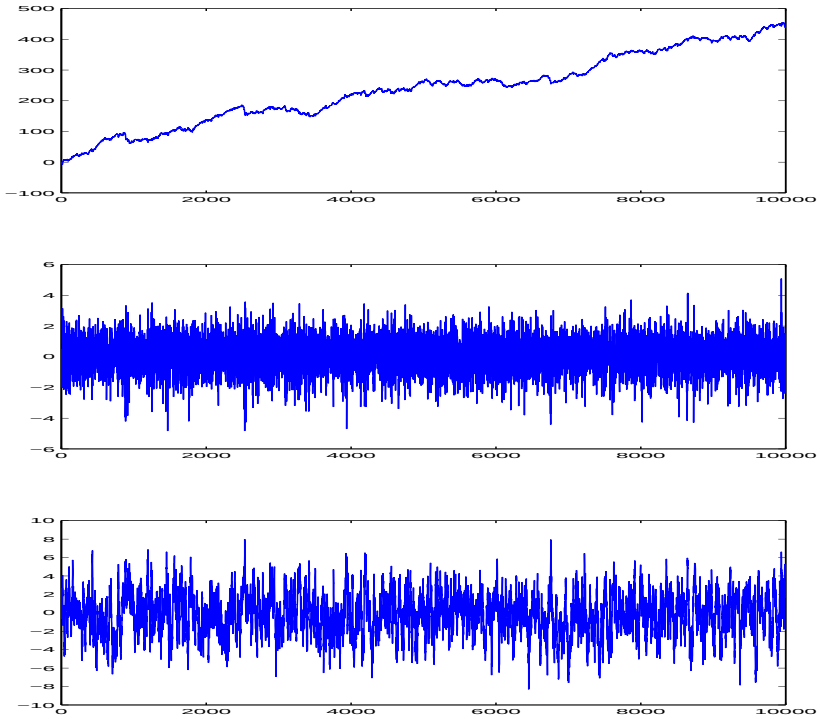

Figure 1: Simulated daily prices, returns and volatility factor respectively from the SV1F model with medium mean reversion

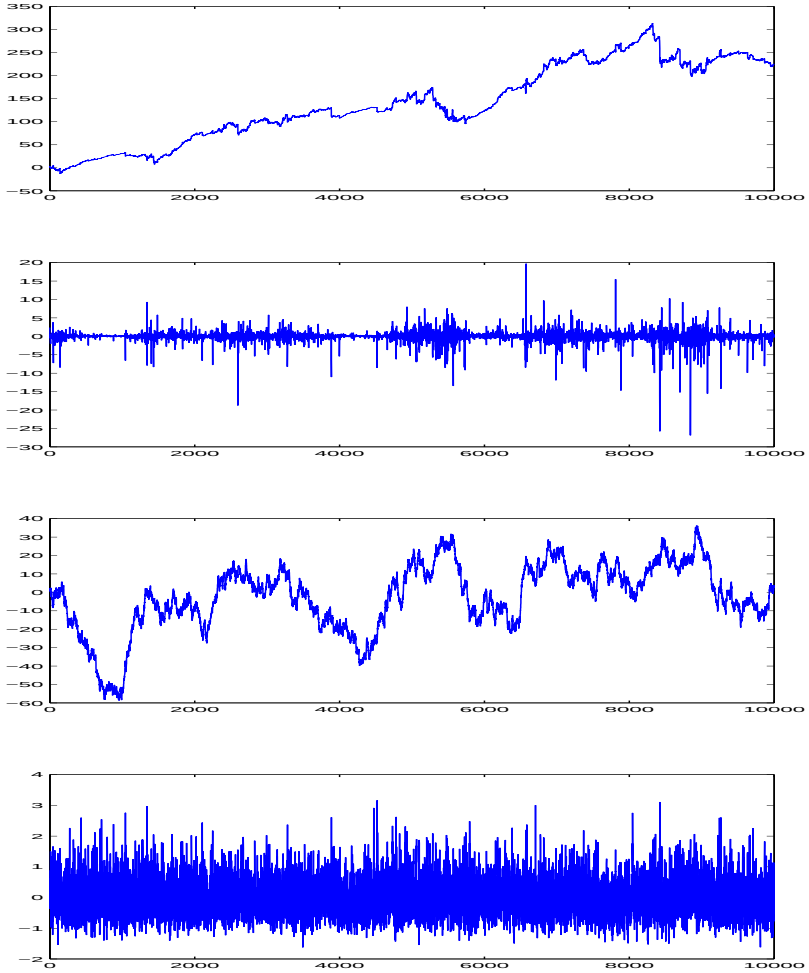

Figure 2: Simulated daily prices, returns and volatility factors respectively from the SV2F model

tends to be oversized at all sampling frequencies. Its size distortion is not very high though, varying around 1-1.5% from the nominal size.

The AJ test statistic was standardized with standard deviations based on both power variations and threshold estimators. In both cases, at a sampling frequency of 1 second, the test seems slightly undersized. However, when diminishing the sampling frequency, the behavior of the test statistics differs. The test becomes rapidly oversized when its variance is based on realized power variations and severely undersized when threshold estimators are used to estimate its variance. This test too seems to work well at higher frequencies.

Table 3 reports the empirical size for the SV2F model.

Procedure	1sec	1 min	5 min	15 min	30 min
AJ(threshold)	0.127	0.094	0.039	0.020	0.008
AJ(power var)	0.052	0.077	0.121	0.205	0.255
BNS	0.054	0.073	0.097	0.113	0.119
CPR	0.062	0.165	0.150	0.168	0.247
JO	0.070	0.106	0.163	0.247	0.327
ABD-LM	0.993	0.699	0.482	0.339	0.254
Med	0.054	0.074	0.102	0.122	0.142
Min	0.052	0.063	0.084	0.082	0.080
PZ	0.701	0.648	0.448	0.305	0.239

Table 3: Size of the tests for jumps for the SV2F model, for a 5% significance level

If we look at all sampling frequencies, the best performance is displayed by the Min test, followed by BNS. For 1 second sampling frequency, size is equal to 5.2% and 5.4%, which increases at lower sampling frequencies though less dramatically than the other tests. The Med, CPR and JO tests behave similar to BNS and Min, but become oversized more rapidly. The AJ(power var) has a size close to the nominal one when sampling is done every second, but then becomes rapidly oversized. When the AJ(threshold) is considered, the test gets severely undersized at lower sampling frequencies. The PZ and the intraday procedures display by far the poorest performance, being severely oversized even when we sample every second (99.3% for the intraday tests and 70.1% for PZ).

POWER We now evaluate the power of the tests by adding to the continuous stochastic volatility process SV1F jump processes with alternative intensities and jump sizes.

Varying jump intensity In order to examine how jump detection changes as the number of jumps grows, we consider Poisson jump arrival times depending on the following varying jump intensities (λ): .014, .058, .089, .118, .5, 1, 1.5, and 2. These intensities can be interpreted as the average number of jumps per day and generate the following total number of jumps: 148, 560, 754, 1208, 5081, 10052, 15058 and 20200. For all these scenarios, we consider a jump size that is normally

distributed with mean 0 and standard deviation equal to 1.5%. We did not impose any restrictions on the maximum number of jumps per day. Thus, more than one jump may occur during a trading day.

In Table 4, we report the size corrected power of the tests by considering some scenarios for the jump intensity. The frequency of correctly identified jumps increases as the jump intensity raises.

λ	Procedure	1sec	1 min	5 min	15 min	30 min
	AJ(threshold)	0.971	0.783	0.223	0.036	0.014
	AJ(power var)	0.970	0.796	0.301	0.183	0.313
	BNS	0.954	0.831	0.702	0.545	0.364
	CPR	0.956	0.851	0.737	0.598	0.449
0.118	JO	0.961	0.831	0.711	0.558	0.408
	ABD-LM	0.984	0.882	0.796	0.673	0.555
	Med	0.950	0.839	0.720	0.582	0.433
	Min	0.939	0.816	0.689	0.510	0.309
	PZ	0.975	0.893	0.779	0.648	0.001
	AJ(threshold)	0.972	0.807	0.232	0.044	0.015
	AJ(power var)	0.972	0.811	0.322	0.216	0.319
	BNS	0.959	0.854	0.728	0.562	0.399
	CPR	0.961	0.870	0.766	0.630	0.504
0.5	JO	0.966	0.853	0.730	0.574	0.445
	ABD-LM	0.985	0.909	0.799	0.663	0.537
	Med	0.955	0.860	0.753	0.603	0.461
	Min	0.949	0.840	0.709	0.544	0.347
	PZ	0.982	0.909	0.804	0.679	0.000
	AJ(threshold)	0.982	0.836	0.224	0.042	0.010
	AJ(power var)	0.982	0.852	0.351	0.224	0.333
	BNS	0.970	0.890	0.782	0.612	0.427
	CPR	0.971	0.905	0.815	0.686	0.538
1	JO	0.975	0.887	0.771	0.612	0.454
	ABD-LM	0.988	0.929	0.840	0.691	0.532
	Med	0.969	0.893	0.795	0.646	0.473
	Min	0.962	0.877	0.759	0.581	0.365
	PZ	0.988	0.930	0.855	0.724	0.000
	AJ(threshold)	0.992	0.858	0.192	0.030	0.009
	AJ(power var)	0.992	0.900	0.409	0.256	0.353
	BNS	0.984	0.933	0.854	0.688	0.485
	CPR	0.986	0.942	0.882	0.778	0.618
2	JO	0.988	0.919	0.823	0.655	0.489
	ABD-LM	0.995	0.957	0.883	0.728	0.533
	Med	0.983	0.930	0.845	0.688	0.515
	Min	0.981	0.924	0.829	0.645	0.420
	PZ	0.994	0.960	0.907	0.800	0.000

Table 4: Size corrected power for varying jump intensities and a 5% significance level.

The best tests in terms of power are the intraday ABD-LM procedures and the PZ test. Let us consider the intraday procedures first. The corrected power for these tests is around 98-99% for a sampling frequency of 1 second and then gradually diminishes as the sampling frequency decreases. As the jump intensity diminishes, the power for these procedures ranges between 88% and 96%, for a sampling frequency of 1 minute, between 79% and 88% for 5 minutes data, between 67% and 73% for 15 minutes and finally between 53% and 55% for 30 minutes.

For the PZ procedure we observe a very high power (around 98% and 99% at 1 sec) which decreases with the sampling frequency. It remains higher than the other procedures (except the

intraday tests) for data sampled at 1, 5 and 15 minutes. It is worth mentioning that at 30 minutes the power of PZ is very close to 0 in all cases, even if the actual power (not reported here) ranges between 50% and 60%. This is due to the fact that the PZ statistic tends to become extremely large at very low frequencies under both the null and the alternative hypotheses. As it can be seen in Table 2,at 30 minutes PZ spuriously detects jumps on 12.1% days. The average of the PZ statistic in this 12.1% cases is $3.29 \cdot 10^{12}$.

The JO test displays a very high power (between 96% and 98%) at 1 second and can be ranked after the PZ, ABD-LM and AJ tests. However, at lower frequencies, its power becomes slightly lower than the other tests, except AJ. Power ranges between 83% and 92% at 1 minute, between 71% and 82% at 5 minutes, between 56% and 66% at 15 minutes and finally between 41% and 49% for data sampled every 30 minutes.

Both versions of the AJ test display a high power at 1 second, which plummets at lower frequencies. For instance, if we look at the results for $\lambda = .5$, the power decreases at around 80% when sampling is done every minute, for both versions of the test, followed by a fall at a level of 23% for the version based on threshold estimators and 32% for the test based on power variations, for a sampling frequency of 5 minutes. If we look at lower frequencies, the test based on power variation-type estimators displays a gradual decrease in power, which gets to a value of 24% for a 30 minutes sampling frequency, while the version based on threshold estimators displays a very low power of 0.6% at 30 minutes.

The BNS, CPR, Med and Min tests display a very similar behaviour. They all exhibit very good power properties, with a power ranging between 95% and 98% when sampling at every second, which then decreases with the decrease in the sampling frequency, with values below the ones observed for the intraday and PZ tests. Generally, over all frequencies, the highest power is displayed by CPR, followed by Med, BNS and Min.

Varying jump size A further insight on the ability of all these procedures to identify jumps can be attained by varying the jump size. In this section, we fix the number of jumps for the entire sample and vary the jump size. However, we maintain its nondeterministic character, by drawing it from a normal distribution with mean 0 and a standard deviation that ranges between 0 and 2 bs with a growth rate of .5. Table 5 reports the power of the jump detection procedures.

Overall, the performance of all tests increases with the size of the jumps. The ranking of the tests is in line with what was found for the case of varying jump intensity.

There is a confirmation about the very good ability of the ABD-LM and PZ tests to detect jumps,

σ_{jump}	Procedure	1sec	1 min	5 min	15 min	30 min
	AJ(threshold)	0.921	0.496	0.108	0.026	0.026
	AJ(power var)	0.921	0.509	0.159	0.120	0.232
	BNS	0.872	0.565	0.340	0.178	0.118
0.5	CPR	0.880	0.615	0.394	0.222	0.146
	JO	0.892	0.566	0.322	0.171	0.123
	ABD-LM	0.964	0.698	0.448	0.245	0.128
	Med	0.865	0.590	0.368	0.208	0.132
	Min	0.843	0.532	0.307	0.147	0.076
	PZ	0.950	0.720	0.482	0.262	0.001
	AJ(threshold)	0.972	0.719	0.202	0.030	0.013
	AJ(power var)	0.972	0.727	0.264	0.171	0.284
	BNS	0.942	0.780	0.611	0.416	0.265
1	CPR	0.947	0.810	0.656	0.483	0.340
	JO	0.956	0.779	0.596	0.418	0.283
	ABD-LM	0.987	0.839	0.687	0.493	0.337
	Med	0.940	0.792	0.637	0.459	0.318
	Min	0.928	0.757	0.588	0.385	0.210
	PZ	0.982	0.865	0.723	0.535	0.000
	AJ(threshold)	0.976	0.802	0.231	0.040	0.016
	AJ(power var)	0.976	0.815	0.331	0.212	0.316
	BNS	0.962	0.861	0.730	0.566	0.401
1.5	CPR	0.965	0.877	0.769	0.626	0.490
	JO	0.968	0.865	0.733	0.572	0.426
	ABD-LM	0.985	0.883	0.771	0.622	0.479
	Med	0.959	0.866	0.751	0.602	0.448
	Min	0.953	0.845	0.713	0.531	0.344
	PZ	0.984	0.912	0.813	0.675	0.001
	AJ(threshold)	0.983	0.850	0.244	0.040	0.011
	AJ(power var)	0.983	0.857	0.376	0.245	0.353
	BNS	0.970	0.891	0.799	0.660	0.501
	CPR	0.973	0.902	0.824	0.708	0.588
2	JO	0.977	0.891	0.794	0.665	0.519
	ABD-LM	0.990	0.901	0.816	0.693	0.569
	Med	0.971	0.892	0.806	0.681	0.544
	Min	0.964	0.880	0.778	0.631	0.447
	PZ	0.988	0.932	0.856	0.744	0.001

Table 5: Size corrected power for a varying jump variance and a 5% significance level.

with power ranging between 95% and 99% at 1 second, which gradually decreases with the sampling frequency. Just as in the case of varying jump intensity, the JO procedure exhibits a very high power at 1 second sampling frequency, ranging between 89% and 98%. However, at lower frequencies, the procedure loses power in front of all other tests with the exception of AJ.

We observe the same ranking as in the previous section for the CPR, Med, BNS and Min procedures. At the highest frequency, they exhibit a power ranging between 84% and 88% for the lowest levels of jump sizes ($\sigma_{jump} = .5$). When σ_{jump} takes its highest value, 2, power is around 97% for all 4 procedures at 1 second. For lower frequencies, the performance of these tests decays.

The AJ does again very well for the highest frequency and ranks itself immediately after the PZ and ABD-LM procedures. However, at lower frequencies, we observe a dramatic decrease in power.

3.2.2 Size and power of the tests in the presence of microstructure noise

The simulation comparison reported so far is based on the assumption that the simulated prices come from continuous time jump diffusion process. However, when we deal with prices of financial assets, this is no longer the case. The observed price process is a discrete one. It is either constant, generating zero returns, or changes a lot from one transaction to another. As a result, transactions impact prices, and market participants may build strategies to exploit the short-term inefficiencies of the market (deviations from a random walk process). There is a vast theoretical and empirical financial literature that tries to understand and exploit these inefficiencies, which are generally denominated microstructure effects. In this paper, we treat these effects as simple noise that obstructs our viewing of the real price process.

Even if the impact of noise on realized variance has been very well documented in the literature, there is not much theoretical work concerning the impact of noise on jump detection. JO find a bias correction for the realized bipower variation in the presence of i.i.d. microstructure noise. Moreover, they show that their test statistic does not diverge in the presence of i.i.d. noise if the number of observations per day is large but remains finite. AJ derive the limit of their test statistic in the presence of i.i.d. noise, as well. They also note that if the distance between observations is small, but not 0, the test statistic does not diverge. PZ prove the validity of the test even in the presence of two types of noise, such as i.i.d. and i.i.d. plus rounding processes.

In what follows, we simulate i.i.d. microstructure noise normally distributed with mean 0 and a varying variance. The noise is then added to the SV1F model with medium mean reversion to study its effects on the statistical properties of the tests for jumps.

SIZE The following values for the standard deviation of the noise were considered: .027, .040, .052, 0.065 and 0.080. Table 6 reports the frequencies of spuriously detected jumps for all tests, under alternative sampling frequencies and noise variances. We only report here results for three values of σ_{noise}, .027, .052 and .080. The full set of results is available upon request.

Apart from the AJ and JO tests, all tests become severely undersized in the presence of microstructure noise with an increasing size distortion as the variance of the noise grows. AJ(threshold) does better than AJ(power var) when lower sampling frequencies are considered. If sampling is made every 15 minutes, the size of AJ(threshold) gets close to the nominal one. When $\sigma_{noise} = 0.052$ (Table 6), size is 3.7% for the version based on threshold estimators, whereas for the other version of the test, it reaches a very high level of 10.9%.

σ_{noise}	Procedure	1sec	1 min	5 min	15 min	30 min
	AJ(threshold)	1.000	0.602	0.062	0.031	0.013
	AJ(power var)	1.000	0.589	0.085	0.095	0.158
	BNS	0.000	0.018	0.051	0.053	0.062
0.027	CPR	0.000	0.025	0.051	0.063	0.077
	JO	0.017	0.035	0.079	0.122	0.188
	ABD-LM	0.013	0.049	0.069	0.065	0.060
	Med	0.000	0.023	0.051	0.055	0.063
	Min	0.000	0.018	0.041	0.037	0.035
	PZ	0.049	0.056	0.086	0.101	0.119
	AJ(threshold)	1.000	0.956	0.160	0.037	0.014
	AJ(power var)	1.000	0.948	0.187	0.109	0.165
	BNS	0.000	0.002	0.043	0.054	0.061
0.052	CPR	0.000	0.003	0.047	0.061	0.075
	JO	0.366	0.017	0.064	0.113	0.185
	ABD-LM	0.009	0.040	0.061	0.059	0.059
	Med	0.000	0.005	0.041	0.055	0.062
	Min	0.000	0.003	0.033	0.034	0.037
	PZ	0.051	0.059	0.087	0.099	0.118
	AJ(threshold)	1.000	0.996	0.304	0.058	0.018
	AJ(power var)	1.000	0.994	0.326	0.148	0.186
	BNS	0.000	0.000	0.025	0.050	0.065
0.08	CPR	0.000	0.000	0.029	0.057	0.072
	JO	0.962	0.011	0.043	0.103	0.179
	ABD-LM	0.011	0.031	0.046	0.055	0.057
	Med	0.000	0.001	0.029	0.051	0.061
	Min	0.000	0.000	0.020	0.033	0.035
	PZ	0.050	0.057	0.074	0.096	0.119

Table 6: Size in the presence of i.i.d. microstructure noise with with varying variance for a 5% significance level

The JO procedure generally displays a very high size in the presence of noise at 1 second, which increases with the variance of the noise. However, when sampling is done at lower frequencies (from 1 minute onward), size decreases abruptly in the beginning and then, moderately increases again. The large size at 1 second is due to the fact that the distribution of the test statistic shifts to the right in the presence of microstructure noise. This effect becomes more intense as the variance of the noise becomes larger. Jiang and Oomen (2008) notice this problem in the original paper and propose corrections for the test statistics in the presence of i.i.d. noise.

The least affected by noise is the PZ procedure, which, at the highest sampling frequency, displays a size close to the nominal one even for the highest values of σ_{noise}. This is a consequence of its higher and rapidly increasing size, which turns out to be an advantage in this case, as it compensates the downward bias caused by the presence of noise. The intraday tests, ABD-LM, also behave very well in the presence of i.i.d. noise, being less underbiased than other procedures at high frequencies.

The BNS, CPR, Med and Min tests are severely undersized at very high frequencies. Then their size increases with the decrease in the sampling frequency. When the noise standard deviation is lowest (.027), BNS, CPR and Med tend to reach a size level close to the nominal one quite soon, at 5 minutes. At lower frequencies, CPR tends to become more oversized than the other two procedures.

When the impact of noise is higher ($\sigma_{noise} = .052$ or $.080$), the three tests manage to reach their nominal size only at 15 minutes. The Min procedure, which tends to be undersized in the absence of noise, displays size levels lower than the nominal one for all frequencies.

Except the PZ test which has a size close to the nominal one at 1 second and 1 minute sampling frequency, as if the noise was not present, all other tests tend to get close to the nominal size as the sampling frequency diminishes: JO somewhere between the 5 and 15 minutes sampling frequencies, AJ, BNS, CPR and Med generally at 15 minutes, and ABD - LM somewhere between 15 and 30 minutes.

POWER In this section we examine how the ability of the tests to detect jumps changes in the presence of microstructure noise. To the process simulated to quantify size in the presence of microstructure noise, we add a jump process with intensity $\lambda = .5$ and jump sizes randomly drawn from a $\mathcal{N}(0, 1.5\%)$. The size adjusted power for all tests and for different scenarios of noise contamination are reported in Table 7.

σ_{noise}	Procedure	1sec	1 min	5 min	15 min	30 min
	AJ(threshold)	0.003	0.142	0.118	0.035	0.013
	AJ(power var)	0.011	0.210	0.225	0.190	0.293
	BNS	0.772	0.828	0.714	0.560	0.395
0.027	CPR	0.791	0.844	0.750	0.628	0.493
	JO	0.7915	0.8284	0.7177	0.5701	0.4254
	ABD-LM	0.927	0.862	0.757	0.616	0.475
	Med	0.766	0.828	0.741	0.602	0.456
	Min	0.735	0.807	0.699	0.533	0.348
	PZ	0.902	0.889	0.805	0.665	0.000
	AJ(threshold)	0.006	0.015	0.036	0.020	0.010
	AJ(power var)	0.019	0.032	0.119	0.161	0.252
	BNS	0.553	0.760	0.686	0.540	0.384
0.052	CPR	0.593	0.786	0.725	0.611	0.484
	JO	0.5570	0.7562	0.6846	0.5547	0.4157
	ABD-LM	0.851	0.820	0.738	0.605	0.466
	Med	0.547	0.773	0.713	0.586	0.444
	Min	0.507	0.740	0.668	0.514	0.344
	PZ	0.809	0.844	0.778	0.656	0.000
	AJ(threshold)	0.009	0.004	0.009	0.011	0.006
	AJ(power var)	0.031	0.019	0.062	0.128	0.230
	BNS	0.357	0.672	0.641	0.505	0.371
0.08	CPR	0.395	0.710	0.681	0.582	0.465
	JO	0.3296	0.6545	0.6343	0.5144	0.3943
	ABD-LM	0.766	0.755	0.700	0.578	0.455
	Med	0.358	0.692	0.664	0.562	0.430
	Min	0.309	0.654	0.620	0.494	0.332
	PZ	0.689	0.776	0.738	0.622	0.000

Table 7: Power of the tests in the presence of i.i.d. microstructure noise with with varying variance for a 5% significance level

The hierarchy of the tests in terms of power remains close to the one for the case of no noise. As the size of the noise standard deviation increases, we observe a decrease in power. The intraday and PZ procedures display again the best power. ABD-LM displays the same tendency of decreasing

power with the decrease in the sampling frequency, as if the noise were not present. For $\sigma_{noise} = .052$ *or* $.08$, PZ seems to be affected by noise at 1 second, but then regains power at 1 minute (84% and 78% respectively). Power at 30 minutes is again extremely low, just as in the case without noise.

BNS, CPR, Med and Min tend to behave similarly again. They suffer a significant loss of power at 1 second, but then tend to regain it. All these tests exhibit a very fast power recovery, occurring at 1 minute. When $\sigma_{noise} = 0.08$, the highest power at 1 minute (71%) is showed by CPR. It is followed by Med, BNS and Min, with closed values for power. Even if this recovery of performance takes place for most tests, power stays lower than in the absence of noise.

JO displays a similar pattern to the above tests, but an overall slightly weaker performance. It tends to be better ranked in comparison with the other procedures for lower levels of noise. There is a decrease in the corrected power at 1 second, followed by a slight recovery of power up to 1 minute or 5 minutes. Power at 1 minute varies between 65% and 83%, depending on the amount of noise. For lower frequencies, power decreases again.

By far the worst performance is observed for the AJ tests, which lose their power at 1 second. For lower frequencies, we notice a slight increase in power. The test based on multipower variations seems to perform somewhat better than the ones based on threshold estimators.

In this section, we observed that in the presence of noise, the size of the various jump detection procedures came close the nominal one when sampling was performed less often. In the case of the power, this effect is much more moderate. Power is only partly regained at 1 minute for almost all tests, in our simulation set-up. At lower frequencies, power tends to decrease, just as when noise is not present.

The results on both size and power show us how the tests statistics behave in the presence of noise. Most tests (except AJ and JO) become severely undersized and they all lose power. However, results on the frequencies at which either size or power are regained are depend on the simulated data generating process, mostly on the type and amount of noise. There is no literature that can help users to select an optimal frequency at which to apply a certain test. Based on our results in Section 5, we generally advice against sampling at frequencies higher than 5 minutes. A rule of thumb in this case could be applying the same procedure at more frequencies and looking at the frequency from which the percentage of detected jumps tends to stabilize.

4 EXTENSIONS TO THE JUMP TESTING PROCEDURES

4.1 Advantages of approximate finite sample distributions for the ABD and LM tests

As already mentioned, the difference between the ABD and LM procedures resides in the choice of the critical values. On one side, we have the Šidák approach for the ABD procedures, which has the advantage of taking into consideration the daily number of observations. On the other side, the LM test makes use of the asymptotic distribution of the maximum and is characterized by simplicity in comparison with the ABD approach.

In this section, we propose an alternative to the above approaches, by making use of simulated critical values for the maximum of the tests statistics. This approach enables us to account for the sample size in the inference process. Moreover, it is shown that it generates higher power than the asymptotic test (LM), accompanied by a manageable size. [We are grateful to Dobrislav Dobrev for suggesting us to explore the use of approximate finite sample distributions.]

According to this procedure, critical values can be obtained in the following way. Let n be the number of daily observations and $\widehat{V}_j$ the local volatility estimate at time t_j, obtained as in Andersen et al. (2007) and Lee and Mykland (2008). At each time, t_j, we simulate a number of n observations from $\mathscr{N}(0, \widehat{V}_j)$ 10,000 times. Thus, we have 10,000 different price paths of n observations each. For every path, we take the maximum over the n observations. The total of 10,000 maximums represent the approximate finite sample from which we select the critical values. Finally, the statistic in (8) is compared to the corresponding critical value selected as above. Just as for the ABD and LM tests, we reject the null of continuity at time t_j, if the test statistic is higher than the critical value.

The proposed approach is based on the so-called "Monte Carlo Reality Check" defined as a simulation-based method for "obtaining a consistent estimate of a p-value for the null in the context of a specification search" (White, 2000, pp 1102).

To assess the performance of our methodology based on simulated critical values, we use data simulated from the SV1F model with medium mean reversion, augmented by jumps and microstructure noise. The latter is sampled from a $\mathscr{N}(0, \sigma_{noise})$, where σ_{noise} takes the same values as in Section 3.2.2. We compare the results in terms of size and power with the ones based on the asymptotic distribution of the maximum (LM test).

The total number of simulated trading days is 10,000, just as in the previous sections. Each day, n intraday observations are made, where n takes different values depending on the sampling frequency, i.e. 1 second, 1, 5, 15 and 30 minutes. This leads to total number of observations equal to $n \cdot 10,000$ and consequently to an equal number of test statistics of the form in (8).

SIZE We quantify size by using three distinct measures. First, for each of the simulated 10,000 trading days, we observe whether the applied procedures rejected the null at least once during that day. We count all days when this occurred and compute its percentage out of the total number of days. We call the resulting indicator 'daily size'. Second, we compute the percentage of rejections of the null out of the total number of observations ($n \cdot 10,000$) and name this second indicator 'overall size'. Finally, the size distortion is computed by subtracting from the overall size the nominal size. Figures 3 and 4 depict all the above measures together with the corresponding nominal sizes for different sampling frequencies for the SV1F model in the presence of i.i.d. noise. We report the results for only two levels of noise variance, 0.052 (medium) and 0.08 (high).

Both figures show that the test based on simulated critical values is significantly less undersized at very high frequencies than the asymptotic procedure (LM). Thus, for both reported values of noise and for all significance levels, size at 1 second is closer to the nominal one than for the LM test. Size for the finite sample adjustment procedure increases then over the nominal one, but remains very close at 1 minute. This indicates that in the presence of i.i.d microstructure noise, this procedure works better than the asymptotic at high frequencies. However, at lower frequencies the procedure tends to become more oversized than its asymptotic counterpart.

Just as Andersen et al. (2007), we recommend the use of low significance levels when applying the finite sample approximations. This leads to higher critical values and consequently to an improved performance.

POWER In order to assess the power of our finite sample adjustment, we add jumps of different intensities to the SV1F model with medium mean reversion. We only report results for $\lambda = 0.5$, $\sigma_{jump} = 1.5\%$ and under contamination with various levels of microstructure noise, as described at the beginning of this section.

We compute both the daily power, as the percentage of days the procedures were able to correctly signal that at least one jump occurred during the day, as well as the overall power, as the proportion of the total observations correctly classified as jumps. The behaviour of these two measures as a function of the sampling frequency is very similar. We only report the daily size adjusted power.

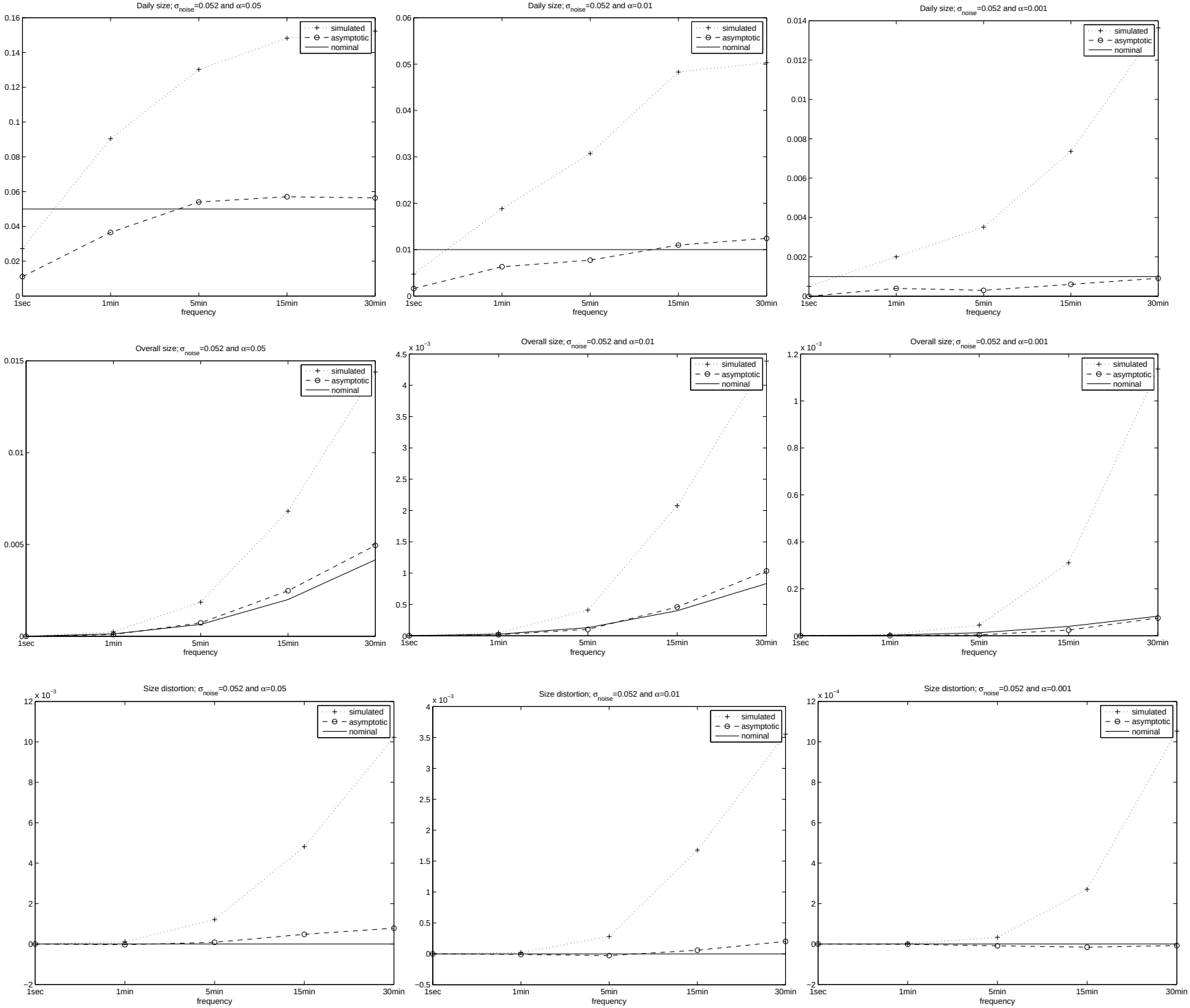

Figure 3: Daily size, overall size and size distortion for simulated and asymptotic critical values based on the SV1F model with noise of variance $\sigma_{noise} = .052$ and for different significance levels: from left to right: 5%, 1%, .1%

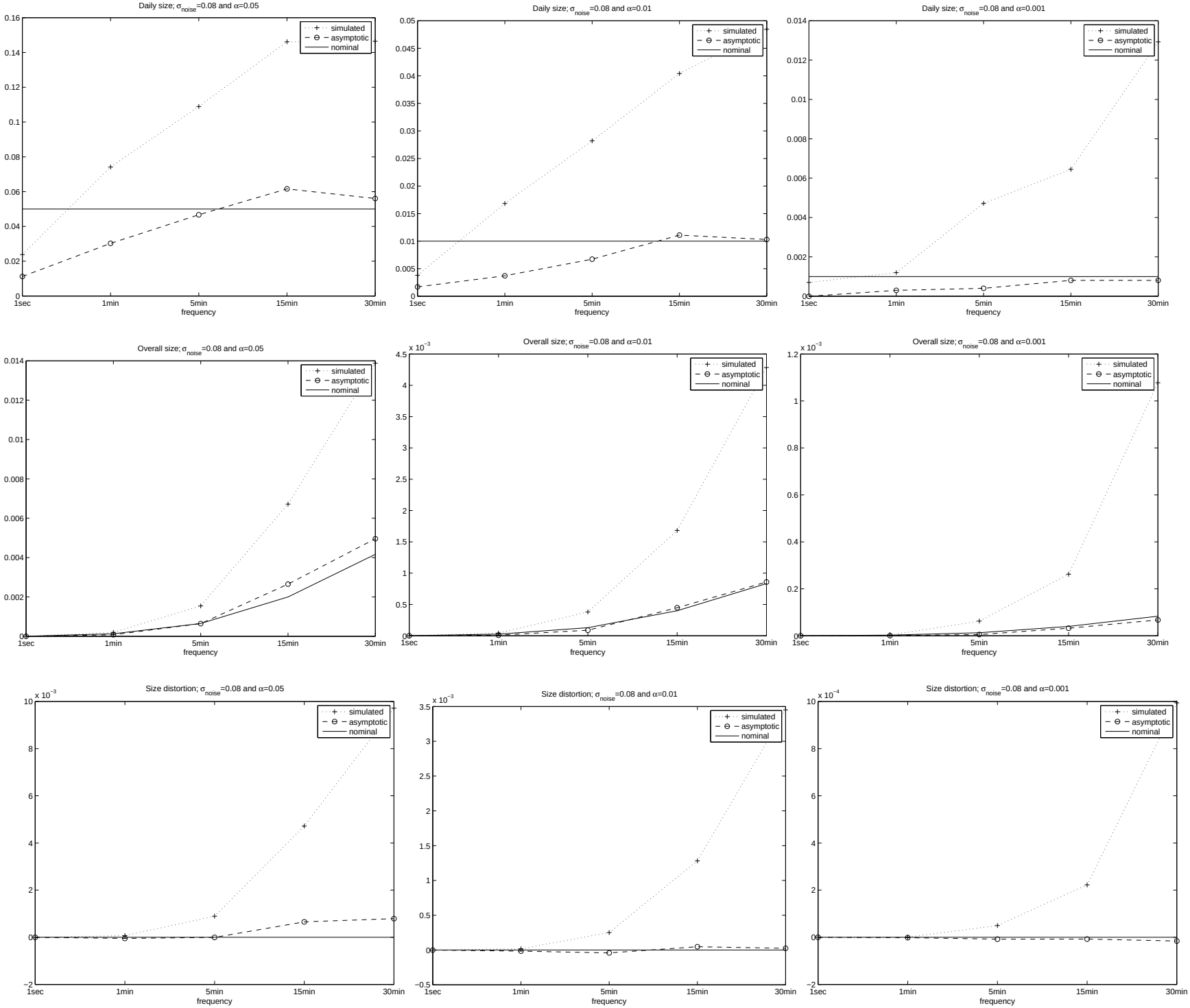

Figure 4: Daily size, overall size and size distortion for simulated and asymptotic critical values based on the SV1F model with noise of variance $\sigma_{noise} = .08$ and for different significance levels: from left to right: 5%, 1%, .1%

Figure 5 illustrates the daily power as a function of the sampling frequency for the three levels of noise variance, 0.027 (low), 0.052 (medium) and 0.08 (high) and different significance levels: 5%, 1% and .1% All the other results for different combinations of jump intensity and noise variance confirm the above results and are not reported but available upon request.

We observe the daily size adjusted power is systematically higher when we use simulated instead of asymptotic critical values over all sampling frequencies and for all significance levels. Moreover, at lower significance levels the gap between the performances of the two approaches seems to widen. To confirm this, we compute power also for significance levels equal to .01% and .001%. For instance, for the case of $\sigma_{noise} = 0.052$ and a sampling frequency of 5 minutes, power at a 5% significance level is 79% for the finite sample adjustment and 76% for the LM test. At a 1% significance level, power for the first procedure is 75%, while power for the second is 73%. At .1%, we have a power of 71% for the first procedure and 68% for the second. At .01%, the first becomes 67%, while second 63%. Finally, at .001%, power for the first is 65%, while for the second 57%.

The main conclusion of this section is that the finite sample adjustment based on simulated critical values leads to a better performance in terms of power and sometimes in terms of size. This approach displays lower size distortions in the presence of microstructure noise at high frequencies. However, at lower frequencies, it tends to become more rapidly oversized than the asymptotic approach. Just as Andersen et al. (2007), we recommend the use of lower significance levels (.1%), which can help to correctly disentangle jumps from the price process, without generating a high number of spurious jumps.

4.1.1 Cross-performances of the tests

This section offers an alternative approach in applying jump tests, which may result quite powerful for empirical purposes. We propose a procedure that combines tests and frequencies suitable in preventing the detection of spurious jumps. We perform this analysis on data simulated based on the SV1F model, augmented by jumps and microstructure noise. Jumps arrive at times sampled from a Poisson distribution with intensity $\lambda = 0.5$ and have a size distributed as a $\mathcal{N}(0, 1.5)$, while the microstructure noise is sampled from a $\mathcal{N}(0, .052)$.

Our simulation analysis revealed that it is worthwhile combining procedures through both reunions and intersections. First, we apply the same procedure at different sampling frequencies, i.e. 1, 5 and 15 minutes. Once the test statistics are computed, we take intersections of the results at 1 and 5 minutes and at 5 and 15 minutes. This leads to two different sets of results. Finally, we

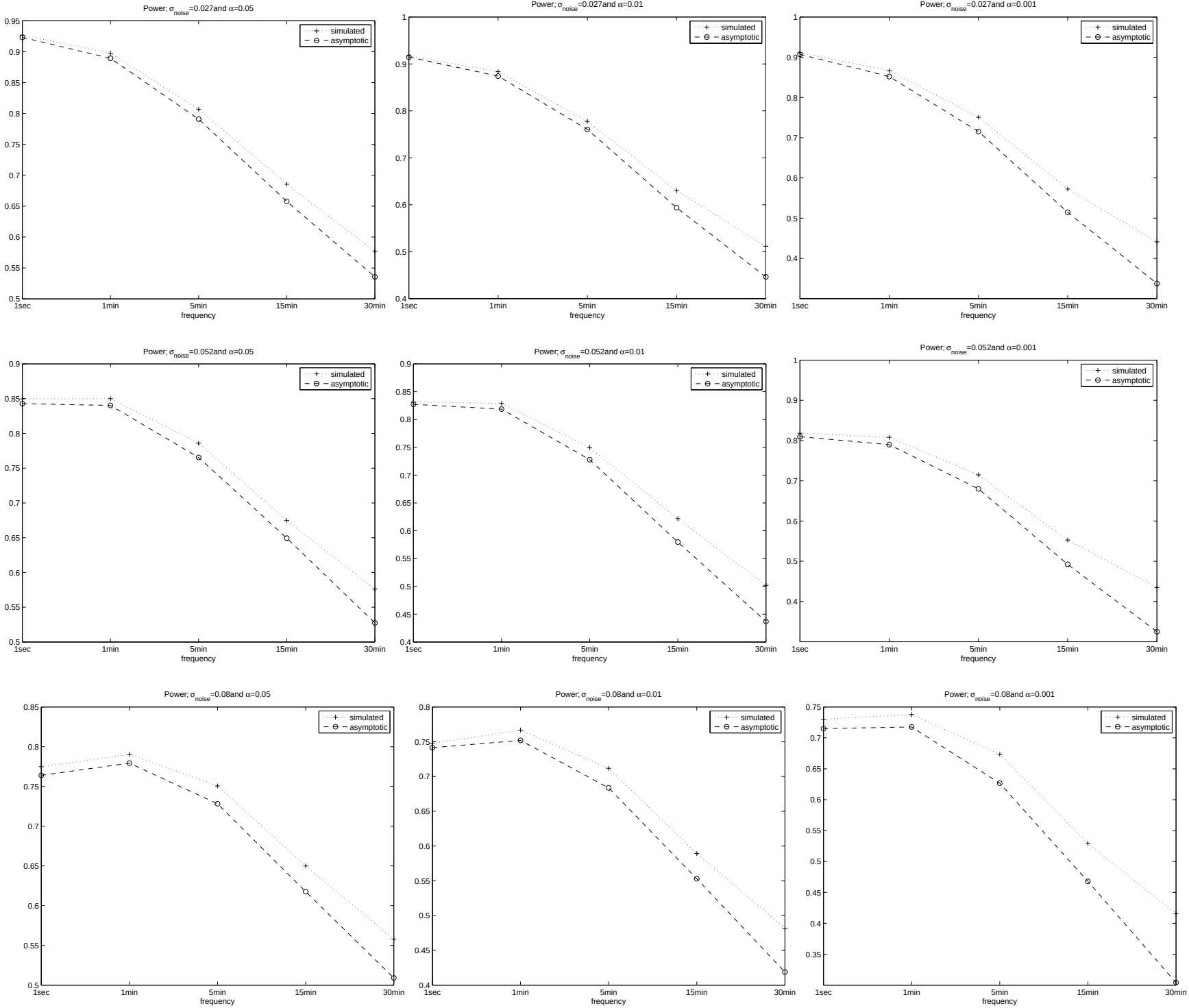

Figure 5: Power for simulated and asymptotic critical values based on the SV1F model with jumps in the presence of noise. Significance levels: 5%, 1%, .1%

take the reunion over the two sets as our final result. For instance, if the considered test is BNS, our decision rule can be written as $(BNS1 \cap BNS5) \cup (BNS5 \cap BNS15)$. This means that on a certain trading day, the path of the price process is considered discontinuous if one or more jumps is/ are detected by the BNS test performed at 5 minutes and at least by one of the other two BNS tests at 1 and 15 minutes.

Table 8 reports the results from combining frequencies for the BNS, CPR, ABD-LM, Med, Min, PZ and JO procedures. In each case, we computed three different measures. First, we report the percentage of correctly classified jumps ('Jump'). Then, we report the percentage of days that are correctly classified as having continuous paths ('No jump'). Finally, we report the percentage of spurious jumps ('Spurious'). The results in Table 8 should be interpreted by contrasting them with the size and power values of the tests reported in Tables 6 and 7. The significance level for all tests is 5%.

Procedure	$(BNS1 \cap BNS5) \cup$ $(BNS5 \cap BNS15)$	$(CPR1 \cap CPR5) \cup$ $(CPR5 \cap CPR15)$	$(ABDLM1 \cap ABDLM5) \cup$ $(ABDLM5 \cap ABDLM15)$	$(Med1 \cap Med5) \cup$ $(Med5 \cap Med15)$	$(Min1 \cap Min5) \cup$ $(Min5 \cap Min15)$
'Jump'	0.6229	0.6772	0.7465	0.6581	0.5953
'No Jump'	0.9574	0.9554	0.9334	0.9583	0.9674
'Spurious'	0.0025	0.0022	0.0247	0.0027	0.0010

Procedure	$(PZ1 \cap PZ5) \cup$ $(PZ5 \cap PZ15)$	$(JO1 \cap JO5) \cup$ $(JO5 \cap JO15)$
'Jump'	0.7744	0.7202
'No Jump'	0.9094	0.9324
'Spurious'	0.0140	0.0206

Table 8: Results from combining tests using different frequencies

The results suggest that our procedure manages to average the power over frequencies and/or tests, combined with a substantial decrease in the percentage of spurious jumps. For instance, in the second column of Table 8, we observe that the percentage of spuriously detected jumps becomes very low (.25%) and is combined with a very high proportion (95.74%) of days that were rightly classified as without jumps and a high proportion of correctly identified jumps (approximately 62.29%). Note that the latter percentage averages out the powers of the BNS test at the given sampling frequencies, i.e. 76% at 1 minute, 69% at 5 minutes, and 54% at 15 minutes (see Table 7).

In Table 8, we notice that one can make the most of this procedure when using a test with a high power, like PZ or ABD-LM. For instance, PZ has a very high power, but also a high size. Combining different frequencies for this test maintains a good power (77%) and at the same time, significantly reduces the percentage of spurious jumps.

In addition to mixing sampling frequencies, we also combine different tests applied on data sampled at the same frequency. Results for some combinations are reported in Table 9 for a sampling frequency of 5 minutes and in Table 10 when sampling is performed every 15 minutes.

'Procedures'	$(Med5 \cap ABDLM5) \cup (ABDLM5 \cap BNS5)$	$(CPR5 \cap BNS5) \cup (BNS5 \cap Med5)$	$(CPR5 \cap BNS5) \cup (BNS5 \cap PZ5)$	$(CPR5 \cap BNS5) \cup (BNS5 \cap Min5)$	$(Med5 \cap BNS5) \cup (BNS5 \cap ABDLM5)$
'Jump'	0.6848	0.6496	0.6658	0.6431	0.6623
'No Jump'	0.9297	0.9434	0.9525	0.9384	0.9543
'Spurious'	0.0119	0.0102	0.0160	0.0084	0.0133
'Procedures'	$(CPR5 \cap ABDLM5) \cup (ABDLM5 \cap PZ5)$	$(JO5 \cap BNS5) \cup (BNS5 \cap PZ5)$	$(BNS5 \cap PZ5) \cup (PZ5 \cap Med5)$	$(CPR5 \cap PZ5) \cup (PZ5 \cap Med5)$	
'Jump'	0.7405	0.6661	0.7165	0.7150	
'No Jump'	0.9298	0.9481	0.9122	0.9028	
'Spurious'	0.0240	0.0158	0.0158	0.0104	

Table 9: Results from combining different tests for jumps for data sampled every 5 minutes

'Procedures'	$(Med15 \cap ABDLM15) \cup (ABDLM15 \cap BNS15)$	$(CPR15 \cap BNS15) \cup (BNS15 \cap Med15)$	$(CPR15 \cap BNS15) \cup (BNS15 \cap PZ15)$	$(CPR15 \cap BNS15) \cup (BNS15 \cap Min15)$	$(Med15 \cap BNS15) \cup (BNS15 \cap ABDLM15)$
'Jump'	0.5465	0.5135	0.5302	0.5067	0.5200
'No Jump'	0.9404	0.9248	0.9359	0.9258	0.9389
'Spurious'	0.0111	0.0217	0.0354	0.0193	0.0171
'Procedures'	$(CPR15 \cap ABDLM15) \cup (ABDLM15 \cap PZ15)$	$(JO15 \cap BNS15) \cup (BNS15 \cap PZ15)$	$(BNS15 \cap PZ15) \cup (PZ15 \cap Med15)$	$(CPR15 \cap PZ15) \cup (PZ15 \cap Med15)$	
'Jump'	0.5984	0.5297	0.5927	0.5962	
'No Jump'	0.9374	0.9265	0.8902	0.8773	
'Spurious'	0.0180	0.0353	0.0359	0.0240	

Table 10: Results from combining different tests for jumps for data sampled every 15 minutes

Just as in the case of combining frequencies, when we combine tests, the percentage of correctly classified jumps ranges between the lowest and the highest powers for individual tests. This effect is accompanied by a significant decrease in the percentage of spurious jumps. From Tables 9 and 10, we observe that the best performance is attained when we use combinations with powerful tests, such as PZ and ABD-LM. Moreover, it is best to intersect one of these tests twice with other procedures. For instance, in Table 9, the sixth combination $((CPR5 \cap ABDLM5) \cup (ABDLM5 \cap PZ5))$ detects all jumps identified by ABD-LM if they are detected by at least one of the CPR and PZ tests. This decision rule generates a high percentage of correctly classified jumps (74%) and a low percentage of spurious jumps (2.4%). The combination $(BNS5 \cap PZ5) \cup (PZ5 \cap Med5)$ intersects PZ twice with two other procedures and manages to attain high power and a low percentage of spurious jumps.

This simple approach is meant to show that combinations of tests and/or sampling frequencies can do better than just applying one single procedure. It preserves a high percentage of rightly classified jumps, with a significant decrease in the percentage of spurious jumps. To maintain a high power, we advise users on combining tests with high power, such as PZ and ABD-LM, with other tests or to combine these tests applied on data sampled at different frequencies.

5 EMPIRICAL APPLICATION

In this final section, we apply all tests for jumps to real financial data. We report an empirical application based on high frequency data for five stocks listed in the New York Stock Exchange, namely Procter&Gamble, IBM, JP Morgan, General Electric and Disney. Our dataset covers 5 years, running from the 3rd of January 2005 to the end of December 2009, with an average of 1250 days.

In order to carry out the jump tests, we rely on transaction data, which we sample at 1, 5, 10, 15 and 30 minutes. This sampling schemes left us with an average of approximately 414 data points at 1 minute, 82 observations at 5 minutes, 40 at 10 minutes, 26 at 15 and 12 at the lowest frequency.

Table 11 reports the proportions of identified jumps.

In general, the proportions of jumps, as well as the behaviour of tests at different frequencies do no vary much from one stock to another. However, for each company, the results obtained from different procedures vary considerably. This reflects once again that these procedures are built in different ways and have very different size and power properties.

For each procedure and for each stock, we observe that there is a decrease in the percentage of identified jumps as we sample less frequent. We can notice this effect better in Figure 6, which includes signature plots of the percentages of identified jumps for all procedures for IBM. Due to the high number of tests, we grouped the procedures. We considered the AJ tests in the first group, while the BNS, CPR, Med and Min made a second group, as they are similarly built. Finally, the rest of the tests, JO, ABD-LM and PZ enter the third group.

At 1 minute, most of the tests detect a high percentage of jumps, which then substantially decreases at 5 minutes. From 5 minutes onward, the decrease in this percentage becomes much slower and a stabilization around 10-15 minutes occurs. We believe that at higher frequencies, the procedures detect a high number of spurious jumps, due to the presence of microstructure noise. A rule of thumb is to apply a test for a variety of frequencies and choose the frequency at which the percentage of jumps stabilizes. In our case, this corresponds to the 10 minutes frequency.

For IBM, PZ and ABD-LM identify 97%, followed by CPR with 88% and BNS with 77%. At lower frequencies, this percentage drastically drops. For instance, the values for the above tests for 1 minute data are 57%(PZ), 51% (ABD-LM), 37%(CPR) and 25% (BNS). This seems contrary to what we observed in Section 3.2.2, where tests statistics are undersized in the presence of microstructure noise.

When tests are based on multipower variations, the reason for the high percentages of detected

Company	Procedure	1 min	5 min	10 min	15 min	30 min
	AJ(threshold)	0.552	0.109	0.050	0.024	0.014
	AJ(power var)	0.606	0.357	0.293	0.264	0.266
	BNS	0.814	0.273	0.154	0.157	0.132
PG	CPR	0.915	0.391	0.221	0.190	0.149
	JO	0.407	0.212	0.188	0.211	0.277
	ABD-LM	0.972	0.506	0.270	0.182	0.086
	Med	0.484	0.174	0.144	0.152	0.140
	Min	0.453	0.157	0.106	0.102	0.074
	PZ	0.969	0.598	0.344	0.278	0.226
	AJ(threshold)	0.534	0.094	0.043	0.020	0.014
	AJ(power var)	0.592	0.330	0.274	0.236	0.237
	BNS	0.765	0.253	0.196	0.191	0.142
IBM	CPR	0.884	0.374	0.257	0.228	0.162
	JO	0.374	0.222	0.230	0.244	0.283
	ABD-LM	0.974	0.512	0.292	0.207	0.097
	Med	0.446	0.174	0.193	0.193	0.156
	Min	0.430	0.148	0.123	0.134	0.090
	PZ	0.967	0.574	0.389	0.325	0.223
	AJ(threshold)	0.548	0.090	0.037	0.031	0.015
	AJ(power var)	0.596	0.317	0.261	0.252	0.263
	BNS	0.708	0.237	0.175	0.155	0.119
JPM	CPR	0.842	0.352	0.237	0.191	0.146
	JO	0.317	0.218	0.212	0.221	0.293
	ABD-LM	0.950	0.500	0.282	0.191	0.121
	Med	0.318	0.167	0.169	0.152	0.132
	Min	0.311	0.134	0.122	0.110	0.065
	PZ	0.953	0.566	0.346	0.269	0.202
	AJ(threshold)	0.563	0.107	0.049	0.034	0.014
	AJ(power var)	0.680	0.368	0.298	0.269	0.310
	BNS	0.754	0.213	0.137	0.153	0.118
GE	CPR	0.908	0.331	0.193	0.196	0.141
	JO	0.317	0.184	0.194	0.199	0.270
	ABD-LM	0.955	0.461	0.259	0.186	0.098
	Med	0.275	0.140	0.128	0.146	0.115
	Min	0.274	0.109	0.092	0.093	0.078
	PZ	0.951	0.510	0.319	0.259	0.194
	AJ(threshold)	0.553	0.086	0.033	0.032	0.016
	AJ(power var)	0.595	0.370	0.323	0.290	0.300
	BNS	0.840	0.327	0.196	0.179	0.135
DIS	CPR	0.923	0.423	0.263	0.217	0.151
	JO	0.385	0.241	0.227	0.246	0.302
	ABD-LM	0.978	0.541	0.271	0.179	0.101
	Med	0.466	0.184	0.188	0.188	0.150
	Min	0.431	0.163	0.118	0.115	0.073
	PZ	0.974	0.595	0.370	0.305	0.209

Table 11: Proportion of days with jumps, at different sampling frequencies, as identified by the following procedures: AJ (both versions), BNS, CPR, JO, ABD-LM, Med, Min and PZ

jumps resides in the fact that data can contain a considerable amount of zero returns when sampled at equal times. As realized multipower variations are computed as the sum of adjacent returns, they tend to be downward biased in the presence of many zero returns. The BNS statistic calculated as the difference between RV_t and BV_t will become bigger as BV_t becomes smaller. The same happens to the ABD-LM statistic, which standardizes returns with BV_t. On the contrary, for tests as Min and Med, based on $MinRV_t$ and $MedRV_t$, this effect is no longer that relevant. We observe that the percentage of detected jumps is 45% and 43% for these tests at 1 minute. The CPR, like BNS is based on a type of multipower variation (threshold), which suffers from the above effect. Moreover, the presence of the threshold makes the multipower variation even smaller, leading to an over-rejection of the null.

The same effect can be noticed for the PZ procedure. The test statistic is based on a threshold volatility estimator, where the threshold is a function of the realized bipower variation. In the presence of zero returns, as BV_t becomes smaller, the threshold also becomes smaller and thus leads to an increase in the test statistic.

The AJ tests display percentages of identified jumps around 55% at 1 second. At lower frequencies, the test based on threshold estimators detects a very small amount of jumps, which is probably due to its lack of power at higher frequencies. On the contrary, the version of the test based on bipower variations tends to identify higher percentages of jumps (between 24% at 30 minutes and 33% at 5 minutes).

The JO test seems only slightly affected by zero returns at 1 minute (37% days with jumps). The percentage of detected jumps does not change very much with the frequency.

The high variability in the percentage of detected jumps reported in Table 11 calls for the application of the combinations of tests as we proposed in Section 4.1.1. Table 12 reports the proportion of jumps detected by different combinations of frequencies (first four lines) and procedures (last 2 lines) for IBM. There is a confirmation that combining procedures leads to a decrease in the proportion of identified jumps. Moreover, there is evidence of higher proportion of jumps when procedures with higher power, like ABDLM and PZ, are combined. When combining frequencies (first four lines), in all cases except the ABD-LM and PZ procedures, the proportion of detected jumps is lower than the proportion identified by the individual procedures on each of the combined frequencies, as reported in Table 11. In the case of the ABD-LM and PZ procedures instead, the combination of frequencies leads to a percentage of jumps in the range of the results obtained on individual procedures, due to the high individual power of the two tests. When combing different tests for jumps for a 10 minutes

sampling frequency (last two lines in Table 12), we observe that the proportion of identified jumps is in the range of the proportions obtained in the case of individual procedures, but it tends to be closer to the lower values for individual procedures. So far, the empirical analysis mostly concerned the percentages of jumps occurring during the period considered.

Procedure	$BNS5 \cap BNS10 \cup$ $BNS10 \cap BNS15$	$(CPR5 \cap CPR10) \cup$ $(CPR10 \cap CPR15)$	$ABDLM5 \cap ABDLM10) \cup$ $(ABDLM10 \cap ABDLM15)$	$(Med5 \cap Med10) \cup$ $(Med10 \cap Med15)$
Proportion	0.105	0.173	0.258	0.098
Procedure	$(Min5 \cap Min10) \cup$ $(Min10 \cap Min15)$	$(PZ5 \cap PZ10) \cup$ $(PZ10 \cap PZ15)$	$(JO5 \cap JO10) \cup$ $(JO10 \cap JO15)$	
Proportion	0.055	0.327	0.148	
Procedure	$(Med10 \cap ABDLM10) \cup$ $(ABDLM10 \cap BNS10)$	$(CPR10 \cap BNS10) \cup$ $(BNS10 \cap Med10)$	$(CPR10 \cap ABDLM10) \cup$ $(ABDLM10 \cap PZ10)$	$(BNS10 \cap PZ10) \cup$ $(PZ10 \cap Med10)$
Proportion	0.132	0.193	0.222	0.213

Table 12: Proportion of jumps identified by different combinations of sampling frequencies and procedures for IBM

Finally, we evaluate the contribution of jumps to the quadratic variation of the price process. For each test for jumps, we detect all days with discontinuities in the price path. Then, we eliminate jumps from prices by removing the highest return in absolute value that occurs on days with jumps. We compute the realized variance on the initial price series sampled every 10 minutes, as well as on the new series without jumps. The first is a proxy for the QV of the price process, whereas the latter for the IV. Table 13, Panel A reports for each test for jumps and for all years considered in our sample, from 2005 to 2009, the estimates of the QV, the IV, as well as of the QV of the jump process for IBM. Panel B reports the same estimates for some combinations of frequencies and procedures. We account for both the levels (first column for each test) and the corresponding percentages (second column for each test).

RV computed on all observations increases from one year to another up to a peak in 2008, when it reaches a level of 0.155. The peak matches the year the sub-prime crisis affected mostly the financial markets. In 2009, RV decreases to 0.058, which is still very high in comparison to tranquil years, such as 2005 and 2006. The levels of RV_C and RV_J vary a lot depending on the jump detection procedure they are based on. Thus, in Table 13, Panel A, during the first two calm years, 2005 and 2006, the percentage of the QV due to jumps is estimated between 8% and 33% by different procedures. However, this percentage is systematically higher in 2006 than 2005 for all tests. During the years of the financial crises, 2007-2009, this percentage drops. A minimum for almost all testing procedures is reached in 2008, the year of maximum volatility, when the percentage of the QV due to jumps varies between 4% and 22%, depending on the procedure. In periods of high volatility, the ability of the tests to pick up jumps is lower, whereas in calmer periods, jumps are much more

Panel A

IBM	Procedure	AJ(threshold)		AJ(power var)		BNS		CPR		JO		ABD-LM		Med		Min		PZ	
Year	Estimator	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%
2005	RV	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0	0.023	100.0
	RV_C	0.017	72.8	0.018	74.9	0.021	90.3	0.019	82.7	0.020	83.5	0.018	79.9	0.021	90.7	0.022	92.4	0.019	79.8
	RV_J	0.006	27.2	0.006	25.1	0.002	9.7	0.004	17.3	0.004	16.5	0.005	20.1	0.002	9.3	0.002	7.6	0.005	20.2
2006	RV	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0	0.025	100.0
	RV_C	0.017	67.1	0.017	68.2	0.020	81.3	0.019	77.4	0.020	81.0	0.019	74.9	0.021	81.9	0.021	83.7	0.019	74.2
	RV_J	0.008	32.9	0.008	31.8	0.005	18.7	0.006	22.6	0.005	19.0	0.006	25.1	0.005	18.1	0.004	16.3	0.006	25.8
2007	RV	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0	0.040	100.0
	RV_C	0.031	77.3	0.032	78.5	0.039	96.1	0.036	88.4	0.036	90.1	0.033	82.3	0.038	94.6	0.039	97.0	0.033	82.0
	RV_J	0.009	22.7	0.009	21.5	0.002	3.9	0.005	11.6	0.004	9.9	0.007	17.7	0.002	5.4	0.001	3.0	0.007	18.0
2008	RV	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0	0.155	100.0
	RV_C	0.121	78.3	0.122	78.4	0.147	94.6	0.140	90.1	0.141	91.2	0.136	87.7	0.145	93.4	0.149	95.9	0.137	88.0
	RV_J	0.034	21.7	0.034	21.6	0.008	5.4	0.015	9.9	0.014	8.8	0.019	12.3	0.010	6.6	0.006	4.1	0.019	12.0
2009	RV	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0	0.058	100.0
	RV_C	0.042	72.6	0.043	74.4	0.052	89.6	0.048	83.9	0.050	86.6	0.047	81.7	0.051	88.3	0.053	92.2	0.046	79.4
	RV_J	0.016	27.4	0.015	25.6	0.006	10.4	0.009	16.1	0.008	13.4	0.011	18.3	0.007	11.7	0.005	7.8	0.012	20.6

Panel B

IBM	Procedure	$BNS5 \cap BNS10 \cup BNS10 \cap BNS15$		$ABDLM5 \cap ABDLM10) \cup (ABDLM10 \cap ABDLM15)$		$(Med5 \cap Med10) \cup (Med10 \cap Med15)$		$(PZ5 \cap PZ10) \cup (PZ10 \cap PZ15)$		$(Med10 \cap ABDLM10) \cup (ABDLM10 \cap BNS10)$		$(CPR10 \cap ABDLM10) \cup (ABDLM10 \cap PZ10)$		$(BNS10 \cap PZ10) \cup (PZ10 \cap Med10)$	
Year	Estimator	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%	Value	%
2005	RV	0.023	100.00	0.023	100.00	0.023	100.00	0.023	100.00	0.023	100.00	0.023	100.00	0.023	100.00
	RV_C	0.022	95.71	0.019	82.19	0.022	93.55	0.019	81.43	0.021	90.51	0.019	81.78	0.021	89.92
	RV_J	0.001	4.29	0.004	17.81	0.002	6.45	18.568	18.57	0.002	9.49	0.004	18.22	0.002	10.08
2006	RV	0.025	100.00	0.025	100.00	0.025	100.00	0.025	100.00	0.025	100.00	0.025	100.00	0.025	100.00
	RV_C	0.022	89.17	0.019	76.15	0.022	88.08	0.019	75.04	0.021	81.89	0.019	77.09	0.020	80.66
	RV_J	0.003	10.84	0.006	23.85	0.003	11.92	24.965	24.97	0.005	18.11	0.006	22.91	0.005	19.34
2007	RV	0.040	100.00	0.040	100.00	0.040	100.00	0.040	100.00	0.040	100.00	0.040	100.00	0.040	100.00
	RV_C	0.039	97.48	0.034	83.03	0.039	97.12	0.033	82.80	0.039	95.67	0.034	83.79	0.038	94.46
	RV_J	0.001	2.52	0.007	16.97	0.001	2.88	17.202	17.20	0.002	4.33	0.007	16.21	0.002	5.54
2008	RV	0.155	100.00	0.155	100.00	0.155	100.00	0.155	100.00	0.155	100.00	0.155	100.00	0.155	100.00
	RV_C	0.151	97.31	0.137	88.38	0.149	96.03	0.138	88.82	0.146	94.27	0.140	90.04	0.144	92.58
	RV_J	0.004	2.69	0.018	11.62	0.006	3.97	11.179	11.18	0.009	5.73	0.015	9.96	0.012	7.42
2009	RV	0.058	100.00	0.058	100.00	0.058	100.00	0.058	100.00	0.058	100.00	0.058	100.00	0.058	100.00
	RV_C	0.053	91.48	0.048	82.51	0.053	91.91	0.047	81.11	0.051	88.77	0.048	83.82	0.050	86.54
	RV_J	0.005	8.52	0.010	17.49	0.005	8.09	18.892	18.89	0.006	11.23	0.009	16.18	0.008	13.47

Table 13: Yearly estimates for the QV of the price (RV), the IV (RV_C), and the QV of the jump process (RV_J), in absolute values and percentages for IBM

visible. Panel B shows the values for RV, RV_C and RV_J for various combinations of frequencies and procedures. As expected, the QV due to jumps is generally lower when combinations are used than when individual tests are applied. When frequencies are combined (first four combinations), RV_J is always lower, whereas when tests are combined, RV_J is in the range of the values for individual tests.

Our results show that tests for jumps produce very different results, both in terms of percentages of identified jumps and the contribution of jumps to the yearly QV. This conclusion supports our proposal to combine tests and sampling frequencies to obtain more clear-cut findings. Consequently, we also perform the empirical analysis for different combinations of frequencies and procedures. This methodology leads to a decrease in the percentage of identified jumps and in the QV due to jumps, which is congruent with the findings in Section 4.1.1, that show that combinations of procedures and frequencies generate fewer spurious jumps.

6 CONCLUSION

The contribution of this paper to the existing literature is twofold. First, we offer a robust and comprehensive comparison between nine alternative jump detection procedures based on high frequency data available in the literature. Second, we offer some useful guidelines to potential users on which test and combinations of tests to use to detect jumps in the prices of financial assets.

To this end, we conducted an extensive numerical analysis using alternative levels of volatility, different levels of persistence in the volatility factor, different jump intensities and jump sizes, different levels of microstructure noise contamination. We also performed an empirical analysis on high frequency data for five stocks listed in the New York Stock Exchange. We summarize the full set of results in Table 14.

It is very difficult to perform a ranking of the tests considering size, power and behaviour in the presence of microstructure noise at the same time. However, for most of the simulated scenarios, the intraday ABD-LM tests for jumps show the best performance. The procedures display a very high power, which is combined with a quite good size behavior. For the stochastic volatility model with one factor, SV1F, size remains relatively stable over different sampling frequencies. The tests also perform very well in the presence of microstructure noise. The use of the intraday tests have the advantage of allowing users to implicitly detect the time and size of the jump. However, they also have two drawbacks. First, in the case of extremely volatile processes, like the stochastic volatility

model with two factors (SV2F), the tests become highly oversized. This is because they standardize each intraday return by a local volatility estimate. When local volatility is very high, the tests will not be able to detect high returns due to jumps. Consequently, their use might not be recommendable for very volatile data. Second, the local volatility of the price process tends to vary a lot during the trading day and exhibits intra-week and intra-day periodicity. The ABD-LM tests do not take into account this factor. Boudt et al. (2009) try to solve this issue by proposing parametric and nonparametric estimators of the periodicity factor that are robust to the presence of jumps.

The PZ test displays high power and a very good behavior in the presence of noise, but is also quite oversized. Its size increases very rapidly when the sampling frequency diminishes. However, given its robustness to microstructure effects, it can be successfully applied at high frequencies, without worrying about the noise.

The BNS, CPR, Med and Min tests display a similar behaviour. They are all built based on comparisons of the realized variation with robust to jumps volatility estimators. They exhibit a size that increases at lower sampling frequencies. CPR tends to be more oversized than the others, whereas Min more undersized. For the SV1F model, BNS and Med can be considered first ranked in terms of size, which remains relatively stable over the varying sampling frequency. BNS has also the most stable size for the SV2F model. All tests also show quite good power. In the presence of microstructure noise, the tests statistics for all these four procedures get very downwards biased and sampling at lower frequencies is obligatory.

The JO test exhibits in the absence of noise a size that is rapidly increasing with the decrease in the sampling frequency. In terms of power, it shows a very high power at very high frequencies, which then decreases at lower frequencies more rapidly than for most of the other tests. In the presence of noise, the test statistics diverges and shifts to the right. Size becomes extremely high at very high frequencies. In addition, the procedure loses power considerably.

There is not a clear-cut behaviour with respect to the AJ procedure. It works well in terms of both size and power only at high frequencies (1 second in our simulation exercise). However, for lower frequencies, there is evidence of a substantial decrease in power, combined with an increase/ decrease in size, depending on how the statistic is computed, multi-power variations/threshold estimators. Moreover, this test becomes extremely oversized at high frequencies in the presence of noise and thus, a very frequent sampling scheme, which could preserve good size and power properties, is not possible.

We applied all jump detection procedures on high frequency data for five stocks listed in the New

York Stock Exchange, namely Procter&Gamble, IBM, JP Morgan, General Electric and Disney, during 2005 and 2009. First, we estimated, for all procedures the percentage of days with jumps for different sampling frequencies, 1, 5, 10, 15 and 30 minutes. We show that the percentage decreases when the sampling frequency diminishes and vary considerably from one procedure to another. Second, we estimated the level and percentage of the yearly quadratic variation coming from the jump process. We show that these estimates differ very much from one procedure to another. In addition, we find that during very volatile years, especially in 2008, the percentage of the quadratic variation caused by jumps reduces in comparison to calm years.

Besides performing a comparison between procedures that identify jumps based on high frequency data, this paper brings two other contributions to the existing literature. First, we propose a finite sample adjustment for the ABD-LM procedure. We suggest computing simulated critical values, as an alternative to the asymptotic critical values. This approach leads to an improvement in the size adjusted power, as well in size at higher frequencies. However, it tends to be more oversized at lower frequencies. In line with Andersen et al. (2007), we recommend the use of lower significance levels (.1%), which can help to correctly disentangle jumps from the price process, without generating a high number of spurious jumps. Second, both the simulation and empirical analyses show that these tests for jumps have different size and power properties and a different behaviour in the presence of market microstructure noise. It is very difficult for users to choose between procedures. Thus, we propose combining these tests through both intersections and reunions over sampling frequencies and procedures. We showed that combining procedures with high power, like PZ or ABD-LM, with other tests leads preserves power, combined with a considerable reduction in the percentage of spurious jumps detected.

The analysis in the present paper can be extended in three different ways. First, for the simulation design, we considered only i.i.d. microstructure noise, in line with most of the papers that introduced these tests to the literature. However, it would be of great interest to observe the impact of zero returns on the behaviour of all these procedures. Second, following the existing literature, in this paper we only considered processes with a finite number of jumps. Thus, a natural extension is a simulation exercise with an infinite number of jumps. Finally, to reduce the probability of detecting spurious jumps, the combination of tests could be enriched by considering test averaging procedures using Fisher (1925)'s method of combining p-values of different tests. We leave these extensions to future research.

Procedure	***Size***	***Power***	***Noise***
AJ (threshold)	slightly undersized; size decreases at lower frequencies	high power at high frequencies which diminishes abruptly at lower frequencies	extremely oversized at very high frequencies, followed by drastic decreases in size from 1 min onward; very high power which decreases abruptly
AJ (power var)	oversized; size rapidly increases across the frequency	high power at high frequencies which diminishes abruptly at lower frequencies	extremely oversized at very high frequencies, followed by drastic decreases in size from 1 min onward; very high power which decreases abruptly
BNS	oversized; size increases slightly across the frequency; stable	high power decreasing gradually; lower numbers than the intraday and 'PZ tests	severely undersized at high frequencies; low power in the presence of noise
CPR	oversized; size increases across the frequency; higher than BNS, Med	high power decreasing gradually; lower numbers than the intraday and 'PZ tests	severely undersized at high frequencies; low power in the presence of noise
JO	oversized; size increases rapidly across the frequency	high power at high frequencies; decreases at lower frequencies	extremely oversized at very high frequencies; low power
ABD-LM	oversized; size varies across the frequency	high power decreasing gradually	undersized in the presence of noise; maintains quite good power properties
Med	oversized; size increases slightly across the frequency; stable	high power decreasing gradually; lower numbers than the intraday and 'PZ tests	severely undersized at high frequencies; low power in the presence of noise
Min	undersized; size decreases slightly across the frequency; stable	power decreasing gradually; lower values than most of the other tests	severely undersized at high frequencies; low power in the presence of noise
PZ	oversized; size increases rapidly across the frequency	high power decreasing gradually	becomes quickly oversized even in the presence of noise; maintains quite good power properties

Table 14: Summary of our results: size and power properties and behavior in the presence of microstructure noise for all the tests

ACKNOWLEDGEMENTS

We are grateful to participants in the Centre of Econometric Analysis Seminar at Cass Business School (15 June 2008), in the 6th Oxmetrics Conference (Cass Business School, 17-18 September 2008), in particular Sir David Hendry, Siem Jan Koopman and Sébastien Laurent, the PhD Workshop in Turin (Polytechnic University of Turin, 20-21 November 2008) for useful comments. George Tauchen provided us with useful suggestions for our simulation design. We are grateful to Mardi Dungey and Abdullah Yalama for their comments and to Brian McGlennon from ICAP, for his help in building and refining our dataset. A special thank to Dobrislav Dobrev for his extremely useful comments and suggestions on a previous version of the paper. We wish to thank the Editor, Jonathan H. Wright, an Associate Editor and two Referees for very useful comments and suggestions which greatly helped to improve the paper. The usual disclaim applies.

References

Aït-Sahalia, Y. (2004), "Disentangling Diffusion from Jumps," *Journal of Financial Economics*, 74, 487–528.

Aït-Sahalia, Y. and Jacod, J. (2008), "Testing for Jumps in a Discretely Observed Process," *Annals of Statistics*, 37, 184–222.

Andersen, T. G., Benzoni, L., and Lund, J. (2002), "An Empirical Investigation of Continuous-Time Equity Return Models," *The Journal of Finance*, 57, 1239–1284.

Andersen, T. G. and Bollerslev, T. (1998), "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International economic review*, 39, 885–905.

Andersen, T. G., Bollerslev, T., and Dobrev, D. (2007), "No-Arbitrage Semi-Martingale Restrictions for Continuous-Time Volatility Models Subject to Leverage Effects, Jumps and I.I.D. Noise: Theory and Testable Distributional Implications," *Journal of Econometrics*, 138, 125–180.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2009), "Jump-Robust Volatility Estimation using Nearest Neighbor Truncation," NBER Working Papers 15533, National Bureau of Economic Research, Inc.

Barndorff-Nielsen, O. and Shephard, N. (2004), "Power and Bipower Variation with Stochastic Volatility and Jumps," *Journal of Financial Econometrics*, 2, 1–48.

— (2006), "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation," *Journal of Financial Econometrics*, 4, 1–30.

Barndorff-Nielsen, O. E., Graversen, S. E., Jacod, J., Podolskij, M., and Shephard, N. (2003), "A Central Limit Theorem for Realised Power and Bipower Variations of Continuous Semimartingales," in *From stochastic analysis to mathematical finance, Festschrift for Albert Shiryaev*, eds. Kabanov, Y. and Lipster, R., Berlin: Springer, vol. 1, pp. 33–68.

Barndorff-Nielsen, O. E., Shephard, N., and Winkel, M. (2006), "Limit Theorems for Multipower Variation in the Presence of Jumps," *Stochastic Processes and their Applications*, 116, 796–806.

Boudt, K., Croux, C., and Laurent, S. (2009), "Robust Estimation of Intraweek Periodicity in Volatility and Jump Detection," Working paper, Faculty of Business and Economics, K.U. Leuven.

Chernov, M., Gallant, A. R., Ghysels, E., and Tauchen, G. (2003), "Alternative Models for Stock Price Dynamics," *Journal of Econometrics*, 116, 225–257.

Corsi, F., Pirino, D., and Renò, R. (2010), "Threshold Bipower Variation and the Impact of Jumps on Volatility Forecasting," *Journal of Econometrics*, 159, 276–288.

Duffie, D., Pan, J., and Singleton, K. (2000), "Transform Analysis and Asset Pricing for Affine Jump-Diffussions," *Econometrica*, 68, 1343–1376.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver and Boyd (Edinburgh).

Gallant, R. A. and Tauchen, G. (2002), "Efficient Method of Moments," Working Paper 02-06, University of North Carolina, Duke University.

Huang, X. and Tauchen, G. (2005), "The Relative Contribution of Jumps to Total Price Variance," *Journal of Financial Econometrics*, 3, 456–499.

Jiang, G. J. and Oomen, R. (2008), "Testing for Jumps when Asset Prices Are Observed with Noise-a "Swap Variance" Approach," *Journal of Econometrics*, 144, 352–370.

Lee, S. S. and Mykland, P. A. (2008), "Jumps in Financial Markets: a New Nonparametric Test and Jump Dynamics," *Review of Financial studies*, 21, 2535–2563.

Mancini, C. (2009), "Non-parametric Threshold Estimation for Models with Stochastic Diffusion Coefficient and Jumps," *Scandinavian Journal of Statistics*, 36, 270–296.

Podolskij, M. and Ziggel, D. (2010), "New Tests for Jumps in Semimartingale Models," *Statistical Inference for Stochastic Processes*, 13, 15–41.

Schwert, M. W. (2009), "Hop, Skip and Jump What Are Modern Jump Tests Finding in Stock Returns?" Working paper, Duke University.

Theodosiou, M. and Žikeš, F. (2010), "A Comprehensive Comparison of Alternative Tests for Jumps in Asset Prices," Working paper, Imperial College London.

White, H. (2000), "A Reality Check For Data Snooping," *Econometrica*, 68, 1097–1126.

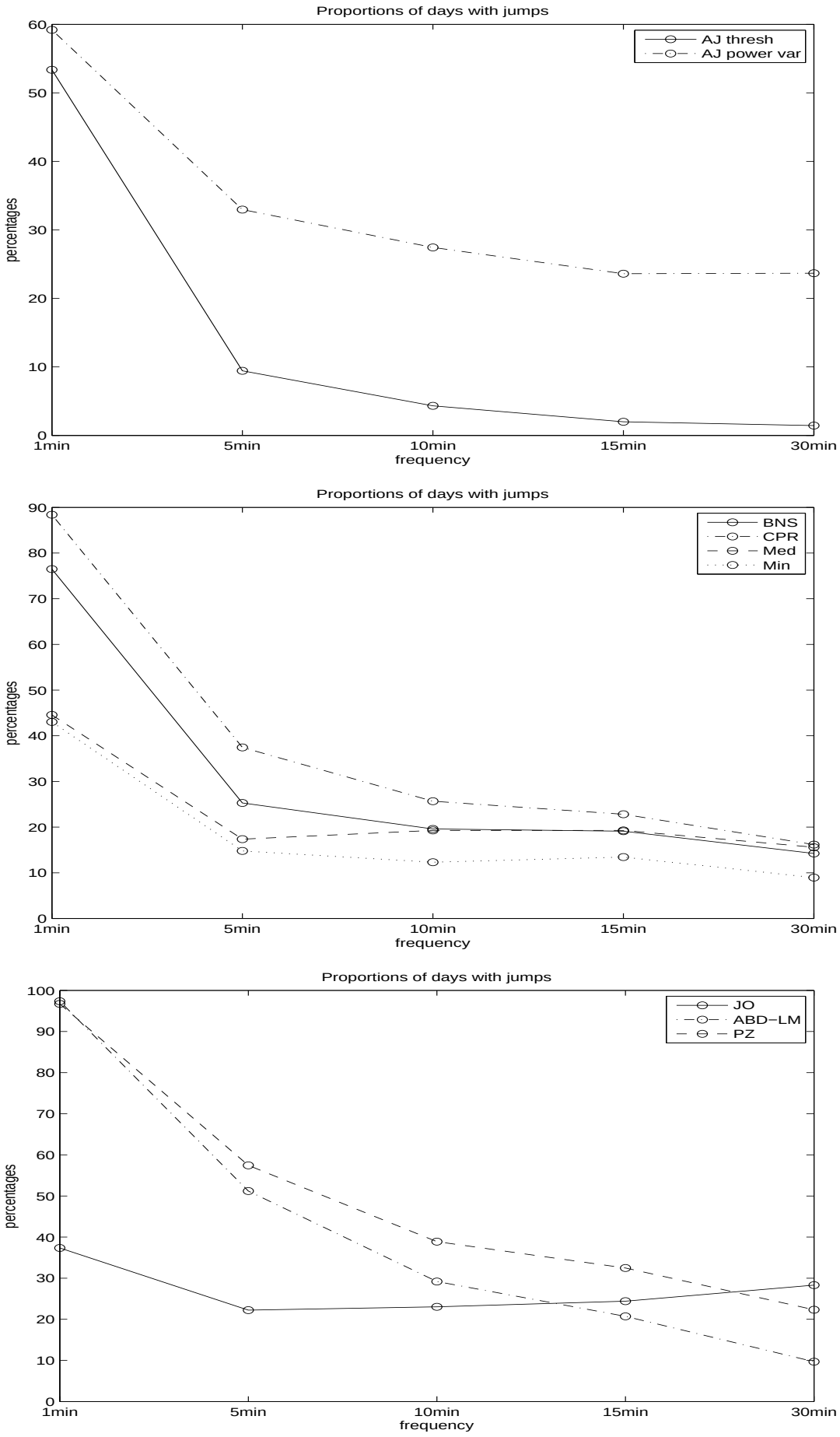

Figure 6: Proportion of days with jumps and sampling frequencies for IBM