Kolmogorov-Smirnov type Tests for Local Gaussianity in High-Frequency Data

George Tauchen, Duke University Viktor Todorov, Northwestern University

July 2013

Abstract

We derive a nonparametric test for the class of Itô semimartingales with non-vanishing diffusion component using high-frequency record of the process on an interval with fixed span. The test is based on the fact that the leading component of the high-frequency increments of Itô semi-martingales with non-vanishing diffusion component is a normally distributed random variable with unknown stochastic variance that is proportional to the length of the high-frequency interval. We form a nonparametric estimate of the local variance and scale the high-frequency increments by it. To remove the effect of "big" jumps, we further discard the high-frequency increments exceeding a time-varying threshold determined by our estimate of the local variance. Our test is then based on comparing the distance between the empirical cdf of the rescaled high-frequency increments not exceeding the threshold and the cdf of a standard normal random variable. We show that the test has a good power against Itô semimartingales with no diffusion component as well as Itô semimartingales contaminated with noise.

Keywords: High-frequency data, Itô semimartingale, jumps, Kolmogorov-Smirnov test, stable process, stochastic volatility

1 Introduction

The standard jump-diffusion model used for modeling many stochastic processes is an Itô semi-martingale given by the following differential equation

$$dX_t = \alpha_t dt + \sigma_t dW_t + dY_t, \tag{1}$$

where α_t and σ_t are processes with càdlàg paths, W_t is a Brownian motion and Y_t is an Itô semimartingale process of pure-jump type (i.e., semimartingale with zero second characteristic, Definition II.2.6 in Jacod and Shiryaev (2003)).

At high-frequencies, provided σ_t does not vanish, the dominant component of X_t is its continuous martingale component and at these frequencies the increments of X_t in (1) behave like scaled and

independent Gaussian random variables. That is, for each fixed t, we have the following convergence

$$\frac{1}{\sqrt{h}}(X_{t+sh} - X_t) \xrightarrow{\mathcal{L}} \sigma_t \times (B_{t+s} - B_s), \quad \text{as } h \to 0 \text{ and } s \in [0,1], \tag{2}$$

where B_t is a Brownian motion and the above convergence is for the Skorokhod topology. There are two distinctive features of the convergence in (2). The first is the scaling factor of the increments on the left side of (2) is the square-root of the length of the high-frequency interval, a feature that has been used in developing tests for presence of diffusion. The second distinctive feature is that the limiting distribution of the (scaled) increments on the right side of (2) is mixed Gaussian (the mixing given by σ_t).

The result in (2) implies that the high-frequency increments are approximately Gaussian but the key obstacle of testing directly (2) is that the variance of the increments, σ_t^2, is unknown and further is approximately constant only over a short interval of time. Therefore, on a first step we split the high-frequency increments into blocks (with length that shrinks asymptotically to zero as we sample more frequently) and form local estimators of volatility over the blocks. We then scale the high-frequency increments within each of the blocks by our local estimates of the volatility. This makes the scaled high-frequency increments approximately i.i.d. centered normal random variables with unit variance. To purge further the effect of "big" jumps, we then discard the increments that exceed a time-varying threshold (that shrinks to zero asymptotically) with time-variation determined by our estimator of the local volatility. We derive a (functional) Central Limit Theorem (CLT) for the convergence of the empirical cdf of the scaled high-frequency increments, not exceeding the threshold, to the cdf of a standard normal random variable. The rate of convergence can be made arbitrary close to $\sqrt{n}$, by appropriately choosing the rate of increase of the block size, where n is the number of high-frequency observations within the time interval. This is achieved despite the use of the block estimators of volatility, each of which can estimate the spot volatility σ_t at a rate no faster than $n^{1/4}$. The developed limit theory then allows us to perform a test of the Kolmogorov-Smirnov type for the local Gaussianity of the high-frequency increments by evaluating maximal difference between two cdf-s.

We further derive the behavior of our statistic in two possible alternatives.

2 Setup

We start with the formal setup and assumptions. We will generalize the setup in (1) to accommodate also the alternative hypothesis in which X can be of pure-jump type. Thus, the generalized setup we consider is the following. The process X is defined on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ and has

the following dynamics

$$dX_t = \alpha_t dt + \sigma_{t-} dS_t + dY_t, \tag{3}$$

where α_t, σ_t and Y_t are processes with càdlàg paths adapted to the filtration and Y_t is of pure-jump type. S_t is a stable process with a characteristic function, see e.g., Sato (1999), given by

$$\log\left[\mathbb{E}(e^{iuS_t})\right] = -t|cu|^\beta \left(1 - i\gamma\text{sign}(u)\Phi\right), \quad \Phi = \begin{cases} \tan(\pi\beta/2) & \text{if } \beta \neq 1, \\ -\frac{2}{\pi}\log|u|, & \text{if } \beta = 1, \end{cases} \tag{4}$$

where $\beta \in (0,2]$ and $\gamma \in [-1,1]$. When $\beta = 2$ and $c = 1/2$ in (4), we recover our original jump-diffusion specification in (1) in the introduction. When $\beta < 2$, X is of pure-jump type. Y_t in (3) will play the role of a "residual" jump component at high frequencies (see assumption A2 below). We note that Y_t can have dependence with S_t (and α_t and σ_t), and thus X_t does not "inherit" the tail properties of the stable process S_t, e.g., X_t can be driven by a tempered stable process whose tail behavior is very different from that of the stable process.

3   Test for Guassianity of High-Frequency Data

We now derive a test for the local Gaussianity at high-frequencies, i.e., the result in (2), using high-frequency record of X. More specifically, we assume X is observed on the equidistant grid $0, \frac{1}{n}, ..., 1$ with $n \to \infty$. We start with the construction of the test statistic followed by limit theory for its behavior under the null and a set of alternatives.

3.1   Empirical CDF of the "Devolatilized" High Frequency Increments

In the derivation of the test statistic we will suppose that S_t is a Brownian motion and then in the next subsection we will derive the behavior of the statistic under the more general case when S_t is stable. The result in (2) suggests that the high-frequency increments $\Delta_i^n X = X_{\frac{i}{n}} - X_{\frac{i-1}{n}}$ are approximately Gaussian with variance given by the value of the process σ_t^2 at the beginning of the increment. Of course, the stochastic volatility σ_t is not known and varies over time. Hence to test for the local Gaussianity of the high-frequency increments we first need to estimate locally σ_t and then divide the high-frequency increments by this estimate. To this end, we divide the interval $[0,1]$ into blocks each of which contains k_n increments, for some deterministic sequence $k_n \to \infty$ with $k_n/n \to 0$. On each of the blocks our local estimator of σ_t^2 is given by

$$\widehat{V}_j^n = \frac{\pi}{2}\frac{n}{k_n} \sum_{i=(j-1)k_n+2}^{jk_n} |\Delta_{i-1}^n X||\Delta_i^n X|, \quad j = 1, ..., \lfloor n/k_n \rfloor. \tag{5}$$

$\widehat{V}_j^n$ is the Bipower Variation proposed by Barndorff-Nielsen and Shephard (2004, 2006) for measuring the quadratic variation of the diffusion component of X. We note that an alternative measure of σ_t can be constructed using the so-called truncated variation. In turns out, however, that while the behavior of the two volatility measures under the null is the same, it differs in the case when S_t is stable with $\beta < 2$. Using a truncated variation type estimator of σ_t will lead to degenerate limit of our statistic, unlike the case of using the Bipower Variation estimator in (5). For this reason we prefer the latter in our analysis.

We use the first m_n increments on each block, with $m_n \leq k_n$, to test for local Gaussianity. The case $m_n = k_n$ amounts to using all increments in the block and we will need $m_n < k_n$ for deriving a feasible CLT later on. Finally, we need to remove the high-frequency increments that contain "big" jumps. The total number of increments used in our statistic is thus given by

$$N^n(\alpha, \varpi) = \sum_{j=1}^{\lfloor n/k_n \rfloor} \sum_{i=(j-1)k_n+1}^{(j-1)k_n+m_n} 1\left(|\Delta_i^n X| \leq \alpha\sqrt{\widehat{V}_j^n}\right), \tag{6}$$

where $\alpha > 0$ an $\varpi \in (0, 1/2)$. We note that we use a time-varying threshold in our truncation to account for the time-varying σ_t. With this, we define

$$\widehat{F}^n(\tau) = \frac{1}{N^n(\alpha, \varpi)} \sum_{j=1}^{\lfloor n/k_n \rfloor} \sum_{i=(j-1)k_n+1}^{(j-1)k_n+m_n} 1\left\{ \frac{\sqrt{n}\Delta_i^n X}{\sqrt{\widehat{V}_j^n}} 1_{\left\{|\Delta_i^n X| \leq \alpha\sqrt{\widehat{V}_j^n}n^{-\varpi}\right\}} \leq \tau \right\}, \tag{7}$$

which is simply the empirical cdf of the "devolatilized" increments that do not contain "big" jumps. Under the null of model (1), $\widehat{F}^n(\tau)$ should be approximately the cdf of a standard normal random variable.

3.2 Convergence in probability of $\widehat{F}^n(\tau)$

We next derive the limiting behavior of $\widehat{F}^n(\tau)$ both under the null of model (1) as well as under a set of alternatives. We start with the case when X_t is given by (3).

Theorem 1 *Under regularity conditions, if the block size grows at the rate*

$$k_n \sim n^q, \quad \text{for some } q \in (0, 1), \tag{8}$$

and $m_n \to \infty$ as $n \to \infty$. Then if $\beta \in (1, 2]$, we have

$$\widehat{F}_n(\tau) \xrightarrow{\mathbb{P}} F_\beta(\tau), \quad \text{as } n \to \infty, \tag{9}$$

where the above convergence is uniform in τ over compact subsets of $(-\infty, 0) \cup (0, +\infty)$; $F_\beta(\tau)$ is the cdf of $\sqrt{\frac{2}{\pi}} \frac{S_1}{\mathbb{E}|S_1|}$ (S_1 is the value of the β-stable process S_t at time 1) and $F_2(\tau)$ equals the cdf of a standard normal variable $\Phi(\tau)$.

We next derive the limiting behavior of $\widehat{F}^n(\tau)$ under an alternative of the local Gaussian setting in (1) which is of particular relevance in financial applications, i.e., the situation when the Itô semimartingale X is "contaminated" by noise.

3.3 CLT for $\widehat{F}_n(\tau)$ under Local Gaussianity

To derive a test based on $\widehat{F}^n(\tau)$ we need a feasible CLT for its behavior under the null of S_t being Brownian motion. This is done in the following theorem.

Theorem 2 *Let X_t satisfy (3) with S_t being a Brownian motion and assume that assumption B holds. Further, let the block size grow at the rate*

$$\frac{m_n}{k_n} \to 0, \quad k_n \sim n^q, \quad \text{for some } q \in (0, 1/2), \text{ when } n \to \infty. \tag{10}$$

We then have locally uniformly in subsets of $(-\infty, 0) \cup (0, +\infty)$

$$\widehat{F}_n(\tau) - \Phi(\tau) = \widehat{Z}_1^n(\tau) + \widehat{Z}_2^n(\tau) + \frac{1}{k_n} \frac{\tau^2 \Phi''(\tau) - \tau \Phi'(\tau)}{8} \left(\left(\frac{\pi}{2}\right)^2 + \pi - 3 \right)$$
$$+ o_p\left(\frac{1}{k_n}\right), \tag{11}$$

$$\left(\sqrt{\lfloor n/k_n \rfloor m_n} \widehat{Z}_1^n(\tau) \quad \sqrt{\lfloor n/k_n \rfloor k_n} \widehat{Z}_2^n(\tau) \right) \overset{\mathcal{L}}{\longrightarrow} (Z_1(\tau) \quad Z_2(\tau)), \tag{12}$$

where $\Phi(\tau)$ is the cdf of a standard normal variable and $Z_1(\tau)$ and $Z_2(\tau)$ are two independent Gaussian processes with covariance functions

$$Cov(Z_1(\tau_1), Z_1(\tau_2)) = \Phi(\tau_1 \wedge \tau_2) - \Phi(\tau_1)\Phi(\tau_2),$$
$$Cov(Z_2(\tau_1), Z_2(\tau_2)) = \left[\frac{\tau_1 \Phi'(\tau_1)}{2} \frac{\tau_2 \Phi'(\tau_2)}{2} \right] \left(\left(\frac{\pi}{2}\right)^2 + \pi - 3 \right), \quad \tau_1, \tau_2 \in \mathbb{R} \setminus 0. \tag{13}$$

3.4 Test for Local Gaussianity

We proceed with a feasible test for a jump-diffusion model of the type given in (1) using Theorem 2.

$$C_n = \left\{ \sup_{\tau \in \mathcal{A}} \sqrt{N^n(\alpha, \varpi)} |\widehat{F}^n(\tau) - \Phi(\tau)| > q_n(\alpha, \mathcal{A}) \right\} \tag{14}$$

where recall $\Phi(\tau)$ denotes the cdf of a standard normal random variable, $\alpha \in (0, 1)$, $\mathcal{A} \in \mathbb{R} \setminus 0$ is a finite union of compact sets with positive Lebesgue measure, and $q_n(\alpha, \mathcal{A})$ is the $(1 - \alpha)$-quantile of

$$\sup_{\tau \in \mathcal{A}} \left| Z_1(\tau) + \sqrt{\frac{m_n}{k_n}} Z_2(\tau) + \sqrt{\frac{m_n}{k_n}} \frac{\sqrt{n}}{k_n} \frac{\tau^2 \Phi''(\tau) - \tau \Phi'(\tau)}{8} \left(\left(\frac{\pi}{2}\right)^2 + \pi - 3 \right) \right|, \tag{15}$$

with $Z_1(\tau)$ and $Z_2(\tau)$ being the Gaussian processes defined in Theorem 2. We can easily evaluate $q_n(\alpha, \mathcal{A})$ via simulation.

Now, in terms of the size and power of the test, under assumptions A and B, using Theorem 1 and Theorem 2, we have

$$\lim_n \mathbb{P}\left(C_n\right) = \alpha, \quad \text{if } \beta = 2 \quad \text{and} \quad \liminf_n \mathbb{P}\left(C_n\right) = 1, \quad \text{if } \beta \in (1, 2), \tag{16}$$

4   Proofs available online

References

Barndorff-Nielsen, O. and N. Shephard (2004). Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics 2*, 1–37.

Barndorff-Nielsen, O. and N. Shephard (2006). Econometrics of Testing for Jumps in Financial Economics using Bipower Variation. *Journal of Financial Econometrics 4*, 1–30.

Jacod, J. and A. Shiryaev (2003). *Limit Theorems For Stochastic Processes* (2nd ed.). Berlin: Springer-Verlag.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge, UK: Cambridge University Press.