M A T H E M A T I C A L E C O N O M I C S

No. 7(14) 2011

THE NORMALITY OF FINANCIAL DATA AFTER AN EXTRACTION OF JUMPS IN THE JUMP-DIFFUSION MODEL

Albert Gardoń

Abstract. When modelling financial data the jump-diffusion processes, driven by Wiener (*W*) and Poisson (*N*) processes, gain increasing importance. On the one hand, they explain better than the Itô diffusion the heavy tails of distributions of percentage changes of stock prices; on the other hand, unlike for example α-stable processes, they are based on the well developed mathematical tools for the Wiener and Poisson processes. After the identification of the jump times, e.g. by means of one of the so-called threshold methods, which are not linked with the continuous part of the model, the parameters from the continuous terms may be estimated similarly as for the Itô diffusion. But it is not obvious if the financial data after an extraction of jumps are already normally distributed. Therefore results of several normality tests will be presented here for chosen data from the Polish stock exchange market.

Keywords: jump-diffusion, Black-Scholes model, threshold method, financial data normality.

JEL Classification: C14, C52.

1. Preliminaries

Let (Ω, F, P) denote the complete probabilistic space and let an Itô process with jumps $X = \{X_t, t \in J = [t_0, T] \subset R_+\}$, whose almost all trajectories are right-hand side continuous left-hand side limits on J (RCLL, càdlàg), be given by the scalar stochastic differential equation in the integral form:

$$X_t = X_{t_0} + \int_{t_0}^{t} a(s, X_s)ds + \int_{t_0^+}^{t} b(s, \overline{X}_s)dW_s + \int_{t_0^+}^{t} c(s, \overline{X}_s)dN_s \,, \qquad (1)$$

Albert Gardoń
Institute of Applied Mathematics, Wrocław University of Economics, Komandorska Street 118/120, 53-345, Wrocław, Poland.
E-mail: albert.gardon@ue.wroc.pl

where $\overline{X}_t = X_{t^-} = \lim_{s \to t-} X_s$, W is a standarized Wiener process, N is a homogeneous Poisson process with intensity λ and both driving processes are said to be independent. Additionally it should be taken into account that all equations and inequalities in this paper hold a.s., i.e. with probability 1, which for the sake of simplicity will not be written down explicitly each time. It is well known (see (Sobczyk, 1991)) that by certain technical assumptions Eqn.(1) has an a.s. (almost surely) unique RCLL-solution. The equation above is also called the jump-diffusion equation and its solution – the jump-diffusion process, and generally it cannot be solved analitycally. Therefore, different numerical methods are developed for approximations of its solution (e.g. see (Gardoń, 2004; Gardoń, 2006)).

The Black-Scholes model describes the behavior of stock prices by means of the special case of Eqn.(1), where the drift $a(t,x) = ax$ and the volatility $b(t,x) = bx$ coefficients are linear, whereas the jump size coefficient $c(t,x) \equiv 0$ vanishes identically (see (Karatzas, Shreve, 1998)), i.e.

$$X_t = X_{t_0} + \int_{t_0}^{t} aX_s\,ds + \int_{t_0}^{t} bX_s\,dW_s\,,$$

whose solution is called an (ordinary) Itô diffusion and has a.s. continuous trajectories. Such a process varies continuously in time. Unfortunately, in practice a stock price may be observed only at chosen (say K) time points, so the model needs to be discretized. Let $(\tau_n)_{n=0}^{\infty}$ be a nondecreasing sequence of stopping times divergent to $+\infty$, then by

$$(\tau)^{\delta} \in \mathsf{P}_{\delta} = \Big\{(\tau_n)_{n=0}^{K} : \tau_0 = t_0, \tau_K = T, \tau_n - \mathsf{F}_{\tau_{n-1}} - measurable, \max_{1\le n\le K}(\tau_n - \tau_{n-1}) = \max_{1\le n\le K}\Delta\tau_n \le \delta\Big\}$$

will be denoted a (random) δ-division of the time interval J. The number δ is called the diameter of the division. This is a very flexible definition allowing a random choice of the observation times, but most often the partition is purely deterministic.

2. The model

The Black-Scholes model requires that the relative price changes:

$$Z_n = \frac{X_{\tau_n} - X_{\tau_{n-1}}}{X_{\tau_{n-1}}} = \frac{\Delta X_{\tau_n}}{X_{\tau_{n-1}}} \approx \ln\frac{X_{\tau_n}}{X_{\tau_{n-1}}}\,, \qquad n = 1,\ldots,K$$

are realizations of normally distributed random variables. But it is well known that this assumption is not valid in practice because of heavy tails of the empirical data distribution (e.g. see (Cont, Tankov, 2004; Johannes, 2004). In other words, relatively large price changes are more probable than follows from the normal distribution assumption. For this reason even more researchers try to model high-frequency financial data by means of alternative stochastic processes. The most popular approach is to treat large price changes as trajectory discontinuities caused by a Poisson process added to the diffusion (e.g. see (Barndorff-Nielsen, Shephard, 2006; Glasserman, Merener, 2003)). In the case of the bond price process, Das (2002) showed that the role of jumps is relevant in incorporating newly released information in interest rate levels, whereas the statistical and economic role of jumps in bond price modeling is further discussed in Johannes (2004). The contribution of the jump component to derivatives pricing is presented, e.g. in Andersen, Bollerslev, Diebold (2007). This manner is more convenient than other models, based on, e.g. α-stable processes because the mathematical tools are very well developed for both the Wiener and Poisson processes. It conducts to a model characterized by the following equation:

$$X_t = X_{t_0} + \int_{t_0}^{t} aX_s\,ds + \int_{t_0^+}^{t} b\overline{X}_s\,dW_s + \int_{t_0^+}^{t} C_s\overline{X}_s\,dN_s\,,$$

where the jump size coefficient C_t is constant in the simplest case, i.e. $C_t \equiv c$, but it may be easy generalized and C_t can denote a strictly stationary process independent from both driving processes (see (Mancini, 2009)). We recall that such a process has the same distribution for each instant $t \in J$ (e.g. Gaussian white noise). In practice this means that the jump sizes are realizations of a random variable of the given distribution.

It is left to estimate coefficients a, b, c and the Poissonian jump rate λ from the data. Since the jump diffusion may be divided into two disjoint parts: the continuous part (ordinary Itô diffusion) and the jump part (discontinuities caused by the Poisson process), the drift a and the volatility b may be evaluated similarly as in the Black-Scholes model after exclusion of the jumps. However the so-called no arbitrage assumption (see (Karatzas, Shreve, 1998)), which excludes the possibility of earning, without any risk more than follows from the risk-free interest rate (say ρ) existing on a market, insists to take $a = \rho$ as the drift. On the contrary, for the ordinary

diffusion the volatility may be estimated in the maximal likelihood sense by the standard deviation of normalized returns:

$$\hat{b} = \sqrt{\frac{1}{K}\sum_{n=1}^{K}\frac{Z_n^2}{\Delta\tau_n} - \left(\frac{1}{K}\sum_{n=1}^{K}\frac{Z_n}{\sqrt{\Delta\tau_n}}\right)^2},$$

which follows immediately from the fact that in the continouos part the relative price changes (Z_n) are normally distributed and b is the infinitesimal variance of the normalized return.

3. Estimators

Jumps could be recognized in several ways. One of them is the nonparametric estimation, proposed previously in Johannes (2004), for which the limiting theory has been fully provided in Bandi, Nguyen (2003). A kind of nonparametric estimation is the so-called threshold method developed by Mancini (2004, 2009). This method consists in the construction of a so-called threshold function, which for each subinterval defines an upper limit for the process change. If the intervals between two observations are getting small, it is possible to distinguish in which intervals the jumps occurred. This is based on the fact that the diffusive part tends to zero at a known rate, namely the modulus of continuity of the Brownian motion paths. This allows identifying asymptotically the jump component and remove it from X in a very effective way. Precisely, under several technical assumptions, jumps should be recognized in the following way:

$$\forall_{n=1,\ldots,K} \qquad \mathbb{I}_{\{\Delta N_{\tau_n}>0\}} = \mathbb{I}_{\{(\Delta X_{\tau_n})^2 > r(\Delta\tau_n)\}},$$

where $\lim_{t\to 0^+} r(t) = 0$ and $\lim_{t\to 0^+} \frac{t\ln t}{r(t)} = 0$. The constraints on r follow from the iterated logarithm low for the standard Brownian motion:

$$\forall_{t\in J} \qquad \limsup_{\delta\to 0^+} \frac{|W_{t+\delta} - W_t|}{\sqrt{-2\delta\ln\delta}} = 1.$$

For this reason, the threshold function should satisfy the evaluation:

$$r(t) > -2t\ln t$$

and usually (see (Mancini, 2009)):

$$r(t) = \beta t^{1-\varepsilon}, \qquad \varepsilon \in \left(0, \frac{1}{2}\right], \quad \beta \geq \frac{2}{e\varepsilon}.$$

Using this bound, the jumps can be extracted and the set of the data should be divided into two parts: the data representing "continuous" price changes and the data representing "discontinuous" price changes. In the second set increments are caused also by the continuous part of the jump-diffusion, but since the diameter of the partition tends to be infinitesimal, these "continuous" price changes are neglected because they are infinitesimal in comparison to the jump sizes. Using this data we can estimate in the maximal likelihood sense:

$$\hat{\lambda} = \frac{N_T - N_{t_0}}{T - t_0},$$

$$\hat{c} = \frac{1}{N_T - N_{t_0}} \sum_{n=1}^{K} Z_n \mathrm{I}_{\{\Delta N_{\tau_n} > 0\}}.$$

The generalization of the jump coefficient to the strictly stationary stochastic process C does not make any additional problem since we have a set of jump realisations of the same distribution. To identify this distribution one can use, e.g. a kernel density estimator.

On the other hand, using the data from the first set, the volatility may be estimated as mentioned in the previous section:

$$\hat{b} = \sqrt{\frac{1}{K - (N_T - N_{t_0})} \sum_{n=1}^{K} \frac{(Z_n^c)^2}{\Delta \tau_n} - \left(\frac{1}{K - (N_T - N_{t_0})} \sum_{n=1}^{K} \frac{Z_n^c}{\sqrt{\Delta \tau_n}} \right)^2}, \qquad (2)$$

where $Z_n^c = Z_n \mathbb{I}_{\{\Delta N_{\tau_n} = 0\}}$. Mancini has also proposed another estimator for the volatility b (see (Mancini, 2009)) based on the quadratic variation of the process X:

$$\hat{b}_M = \sqrt{\frac{\sum_{n=1}^{K} (\Delta X_{\tau_n})^2 \mathrm{I}_{\{\Delta N_{\tau_n} = 0\}}}{\sum_{n=1}^{K} X_{\tau_{n-1}}^2 \Delta \tau_n}}, \qquad (3)$$

which is also consistent. Moreover, for this estimator the speed of convergence with probability 1 is known and equal $\frac{1}{2}$. Besides, the complete results for estimation of the continuous and jump part with quasi MLE method one can find in Ogihara, Yoshida (2011) and Shimizu, Yoshida (2006).

Unfortunately, it turns out that the proposed method does not identify jumps properly, especially in the cases when the stock prices vary strongly (see (Gardoń, 2010)). In such cases either only a few or almost all data are recognized as jumps. The possible cause of such a situation may be the fact that, as it was mentioned before, the construction of the threshold $(\Delta X_{\tau_n})^2 > r(\Delta\tau_n)$ is based on the iterated logarithm law for the standardized Brownian motion. But the continuous part of the process X is a geometrical Brownian motion for which only the inequality below holds (see (Karatzas, Shreve, 1998)):

$$\sup_n \frac{\left| \int_{\tau_{n-1}}^{\tau_n} aX_t\,dt + \int_{\tau_{n-1}}^{\tau_n} bX_t\,dW_t \right|}{\sqrt{-2\delta\ln\delta}} \le M < \infty,$$

where the left-hand side is bounded by a finite random variable M instead of the constant 1. From the theoretical point of view, all is correct because Mancini's proposition is a limit theorem and constants do not play any essential role when δ becomes infinitesimal. But in practice the frequency of observation is bounded from below. Mostly it is a day, a minute or a second. Therefore, the threshold condition should be modified in order to work properly with the real data (see (Gardoń, 2010)):

$$\frac{Z_n^2}{\hat{b}^2} > r(\Delta\tau_n). \tag{4}$$

The method based on the threshold above identifies jumps more efficiently than the original criterion; however, it requires the knowledge of $\hat{b}$ before the extraction of jumps. Nonetheless this problem has been also solved in the cited article by means of an iterative procedure.

4. The normality of a continuous part

The main question we try to answer is if the relative price changes Z_n, after such an extraction of jumps, are already normally distributed or if there are also other problems with the normality, not only heavy tails of the empirical distribution. But in order to check the normality the data needs firstly to be well prepared. Time spans between consequtive observations are not equidistance because of possible jumps (such observations will be excluded from the continuous part of the data) and trading breaks as nights, weekends or holidays, or when the market is closed. Thus, the relative price changes could not be treated as a realization of a sequence of independent, identically distributed random variables although they are indeed independent; the only problem is the distributions identity.

For an infinitesimal time step we have:

$$Z_n \approx \rho(\tau_n - \tau_{n-1}) + b\left(W_{\tau_n} - W_{\tau_{n-1}}\right),$$

which implies:

$$\forall n = 1,\ldots,K-(N_T - N_{t_0}) \qquad Z_n^* = \frac{Z_n - \rho(\tau_n - \tau_{n-1})}{b\sqrt{\tau_n - \tau_{n-1}}} \overset{A}{\sim} \mathrm{N}(0,1)\,.$$

Now the standardized relative price changes (Z_n^*) build the iid $\sim N(0,1)$ – sequence which the normality tests can be conducted for. We have done them for the real financial data from GPW[1] in Warsaw, Poland. As an asset we have chosen The Polish Stock Exchange Index WIG,[2] whose price process is presented in Figure 1, whereas the distribution of the standardized relative price changes before the extraction of jumps *versus* the standard normal curve is shown in Figure 2. The high frequency data from over 10 years consists of over one million ($K = 1'095'645$) each trading minute closing prices from 17 November 2000 to 16 August 2011. The choice of a stock exchange index was not random. Due to the idea of Lapunov's Central Limit Theorem, the phenomena affected by a large number of components with a similar influence on a joint result should be approximately normally distributed. Hence, a stock exchange index, as a kind of an average, should be the best potential candidate for the normality. Its value is affected by prices of almost 400 assets, though their similar influence could be

[1] GPW (*Giełda Papierów Wartościowych*) – stock exchange market.
[2] WIG (*Warszawski Indeks Giełdowy*) – literally: "Warsaw Stock Exchange Index".

questionable. However, due to the construction of the index no component may exceed the level of 10% of influence and no market sector may exceed the level of 30% of influence on the joint value.

Looking at the figures there are two immediatelly visible problems with the normality of this empirical data. The first one, which should be solved by an identification and extraction of jumps, is heavy tails. The second one is the enormous number of small negative tics. It turned out that this was caused by such trading minutes during which the price did not change at all. Theoretically, if the price was to behave as a certain modification of the Brownian motion it should not be constant in any time interval. But in practice we observe only closing prices at the end of each given time interval, the price does not change strictly continuously in time, only if a trading offer occurs and at last the price is rounded. Since 1 October 2002 the WIG value has been rounded to 0.01 point and earlier it was rounded even to integer. For this reason, we have decided to omit such "null tics" and this has helped in a visible way which one can see in Figure 3.

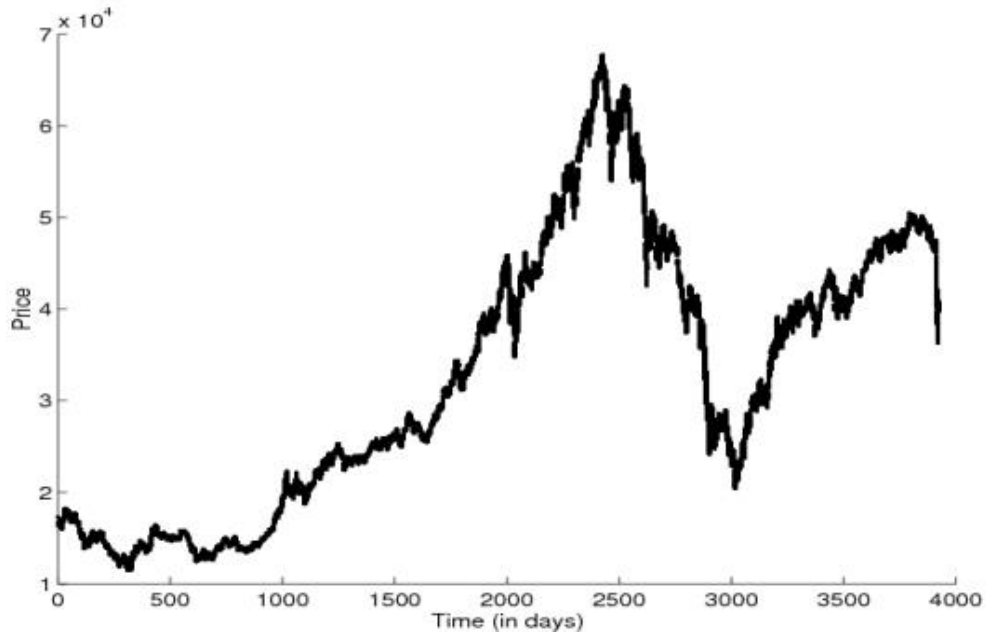

Fig. 1. The price process X (each trading minute closing price) of the Polish Stock Exchange Index (WIG) from 17 November 2000 to 16 August 2011 (3925 days)

Source: author's own study.

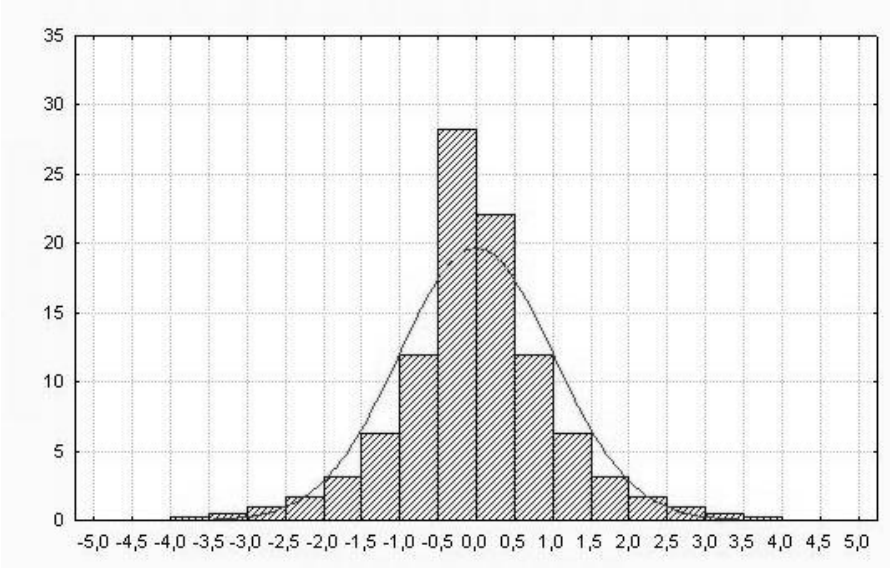

Fig. 2. The distribution of the standardized relative price changes of the Polish Stock Exchange Index (WIG) from 17 November 2000 to 16 August 2011 ($K = 1'095'645$ tics) *versus* the standard normal curve

Source: author's own study.

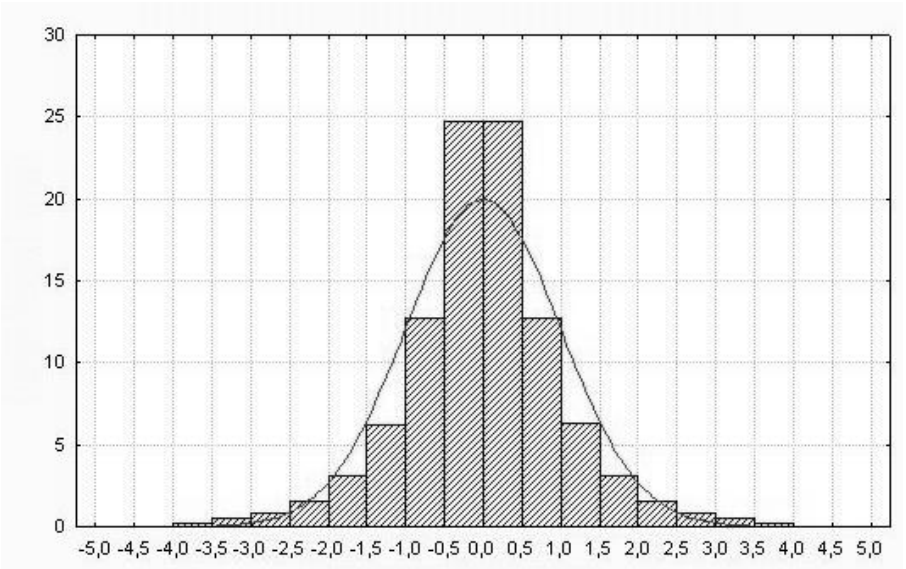

Fig. 3. The distribution of the standardized relative price changes of the Polish Stock Exchange Index (WIG) from 17 November 2000 to 16 August 2011 without "null tics" *versus* the standard normal curve

Source: author's own study.

Table 1. The number of jumps recognized with respect to the given time unit and the given frequency

N_T	Frequency (min)	1	5	15	30	60	1440
U	15 minutes	39 722	11 050	4 689	3 062	1 942	421
N	1 hour	28 927	8 006	3 459	2 191	1 358	282
I	1 day	14 222	3 873	1 638	992	605	128
T	1 week	9 186	2 526	1 082	609	383	90

Source: author's own study.

Table 2. The estimated value of the volatility *b* with respect to the given time unit and the given frequency.

$\hat{b}$	Frequency (min)	1–60	1440
U	15 minutes	0.011–0.013	0.0008
N	1 hour	0.025–0.027	0.018
I	1 day	0.0140–0.0142	0.105
T	1 week	0.0386–0.0396	0.0292

Source: author's own study.

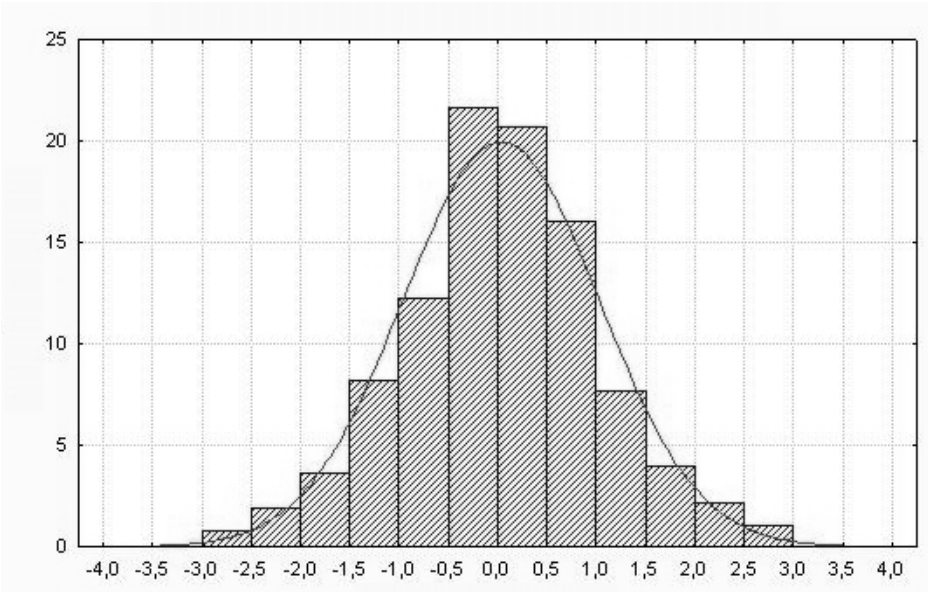

Fig. 4. The distribution of the standardized daily returns (time unit 1 week) of the Polish Stock Exchange Index (WIG) *versus* the standard normal curve

Source: author's own study.

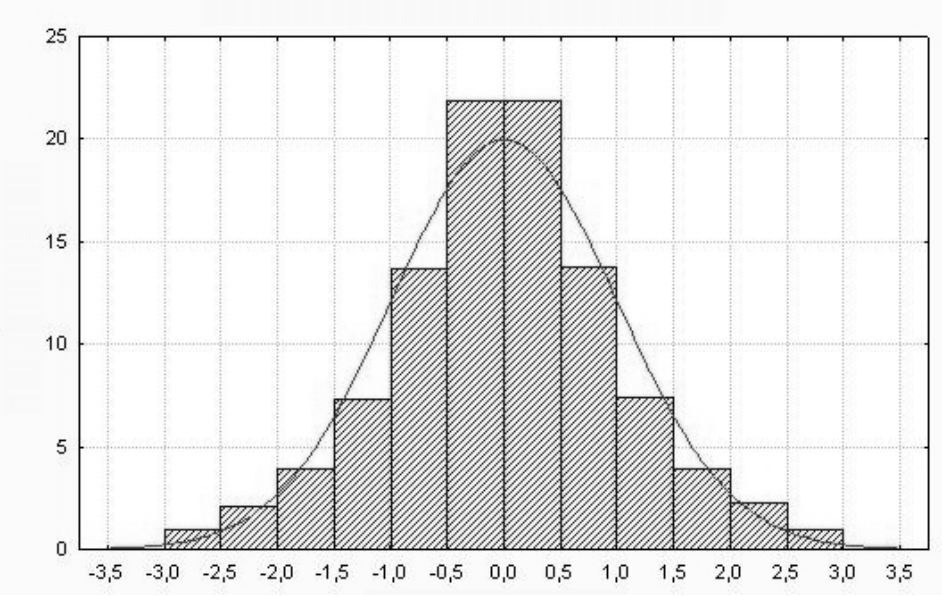

Fig. 5. The distribution symmetry of the standardized high frequency returns (unit: 15 minutes, frequency: 5 minutes) of the Polish Stock Exchange Index (WIG)

Source: author's own study.

The following parameters are fixed in the model: for the sake of simplicity the start time $t_0 = 0$, the end time T = 3924.42 days (which is equivalent to 5'651'165 minutes) and the riskless interest rate ρ = 4.5% yearly (which is equivalent to $1.23 \cdot 10^{-4}$ daily, $8.56 \cdot 10^{-8}$ per minute). For the threshold function r we have chosen $\varepsilon = 0{,}1 \in (0, \frac{1}{2}]$ and $\beta = \frac{2}{e\varepsilon} = 7.3576$ which yields:

$$r(t) = \beta t^{1-\varepsilon} = 7.3576 t^{0.9}.$$

The choice of the unit could not be random. Firstly, since the background on which this application is based is a limit theorem and the division is not equidistance, the unit must be chosen in such a way that for both minimal and maximal time spans between two consecutive partition points the threshold function $r(t)$ is similar as the iterated logarithm bound $-2t \ln t$. Further, we have observed that the number of jumps found and thus the

realizations of Z_n^* apart from jump times, depend on both the data frequency and the time unit assumed. For different time units and different observation frequences (1440 minutes are equivalent to 1 day), we obtained different numbers of jumps which one can see in Table 1. Consequently they have given different estimates for the volatility b, which can be found in Table 2. The normality tests (chi-squared, Lilieforse and Shapiro-Wilk) were conducted for all the combinations of unit and frequency mentioned in those tables. The best choice of their combination, in the sense of the maximal *p*-value, was the time unit 1 week and the data frequency 1 day (1440 minutes). The histogram of the empirical returns distribution in that instance is shown in Figure 4. But unfortunately, even in this best case, all normality tests have suggested the rejection of the normality null hypothesis. The chi-squared test statistic empirical value with 9 degrees of freedom has been equal to 50.3, which corresponds with the empirical significance level $p < 10^{-5}$...

5. Conclusions

Even in the case of the best possible candidate for the normality of the returns distribution, namely a stock exchange index, independently from the chosen time unit and observation frequency, all normality tests insist strongly on rejecting the null hypothesis. The empirical distribution after an exclusion of jumps seems sometimes virtually similar to the normal one, though the very large sample size requires much greater goodness of fit in the statistical sense to the distribution assumed. It turns out that heavy tails of the empirical distribution are not the only problem. This distribution seems to be also too slender for normality; in other words, it has a positive *kurtosis* or the sample is more concentrated about the central point of the distribution than in the normal case.

Our experiments have shown additionally another aspect of returns distribution. It is a very popular statement (e.g. see (Peiró, 1999)) that the financial data is left-hand side skewed. But our research contradicts this opinion. After the elimination of those sample times for which the price process did not change and for which a jump has been detected, the empirical distribution is pretty symmetrical (see Figure 5), at least in the case of the aforementioned index. This means that the absolute decreases could appear statistically greater than the absolute increases because the price process falls from higher levels and rises from lower ones, but percentage decreases and increases seem to be similar.

Literature

Andersen T., Bollerslev T., Diebold F. (2007). *Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility*. Review of Economics and Statistics. Vol. 89. Pp. 701-720.

Bandi F., Nguyen T. (2003). *On the functional estimation of jump-diffusion models*. Journal of Econometrics. Vol. 116. Pp. 293-328.

Barndorff-Nielsen O.E., Shephard N. (2006). *Econometrics of testing for jumps in financial economics using bipower variation*. Journal of Financial Econometrics. Vol. 4. Pp. 1-30.

Cont R., Tankov P. (2004). *Financial Modelling with Jump Processes.* Chapman & Hall – CRC.

Das S. (2002). *The surprise element: Jumps in interest rates*. Journal of Econometrics. Vol. 106. Pp. 27-65.

Gardoń A. (2004). *The order of approximations for solutions of it ô-type stochastic differential equations with jumps*. Stochastic Analysis and Applications. Vol. 22. No. 3. Pp. 679-699.

Gardoń A. (2006). *The order 1.5 approximations for solutions of jump-diffusion equations*. Stochastic Analysis and Applications. Vol. 24. No. 6. Pp. 1147--1168.

Gardoń A. (2010). *The identification of discontinuities for the jump-diffusion process by means of a modified threshold method*. In: *Proceedings of the International Scientific Conference AMSE 2010*, Demänovská Dolina, Slovakia, 26-29 August 2010, Pp. 105-114

Glasserman P., Merener M. (2003). *Numerical solution of jump-diffusion LIBOR market models*. Finance and Stochastics. Vol. 7. Pp. 1-27.

Johannes M. (2004). *The statistical and economic role of jumps in continuous-time interest rate models*. Journal of Finance. Vol. 59. Pp. 227-260.

Karatzas I., Shreve S.E. (1998). *Methods of Mathematical Finance*. Springer--Verlag. New York.

Kloeden P.E., Platen E. (1995). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag. New York–Berlin–Heidelberg.

Mancini C. (2004). *Estimation of the parameters of jump of a general poisson-diffusion model*. Scandinavian Actuarial Journal. Vol. 1. Pp. 42-52.

Mancini C. (2009). *Non parametric threshold estimation for models with stochastic diffusion coefficient and jumps*. Scandinavian Journal of Statistics. Vol. 36. Issue 2. Pp. 270-296.

Ogihara T., Yoshida N. (2011). *Quasi-likelihood analysis for the stochastic differential equation with jumps*. Statistical Inference for Stochastic Processes. Vol. 14. Pp. 189-229.

Peiró A. (1999). *Skewness in financial returns*. Journal of Banking and Finance. Vol. 23. Issue 6. Pp. 847-862.

Shimizu Y., Yoshida N. (2006). *Estimation of parameters for diffusion processes with jumps from discrete observations*. Statistical Inference for Stochastic Processes. Vol. 9. Pp. 227-277.

Sobczyk K. (1991). *Stochastic Differential Equations with Applications to Physics and Engineering*. Kluwer Academic Publishers B.V. Dordrecht.