Nonidentically distributed variables and nonlinear autocorrelation

Annibal Figueiredo[a,e], **Iram Gleria**[b,e], **Raul Matsushita**[c], **Sergio Da Silva**[d,e*]

[a]*Department of Physics, University of Brasilia, 70910-900 Brasilia DF, Brazil*

[b]*Department of Physics, Federal University of Alagoas, 57072-970 Maceio AL, Brazil*

[c]*Department of Statistics, University of Brasilia, 70910-900 Brasilia DF, Brazil*

[d]*Department of Economics, Federal University of Santa Catarina, 88049-970 Florianopolis SC, Brazil*

[e]*National Council for Scientific and Technological Development, Brazil*

Abstract

This paper considers independently distributed stochastic processes that are also nonidentically distributed. We find that an identically distributed process with autocorrelations can be obtained from an independent, yet nonidentically distributed, random generator. Our approach is illustrated with a time series from the British pound-US dollar rate.

PACS: 05.40.+j;02.50.-r

Keywords: Central limit theorem; Independent and identically distributed variables; Autocorrelation

* Corresponding author.
E-mail address: email@sergiodasilva.com (S. Da Silva).

1. Introduction

The aim of this paper is to show how a stochastic time series, obtained from a random generator that is independent but not identically distributed, shows nonlinear autocorrelations when approached as if it were identically distributed. We put forward reduced variables that are independent and identically distributed and observe that the main features of nonlinear autocorrelations emerge. We illustrate our approach with a time series from the British pound-US dollar rate.

The structure of the paper is as follows. Section 2 presents benchmark definitions. The set of independent and identically distributed, reduced variables is presented in Section 3. Section 4 calibrates our methodology with data from the pound-dollar exchange rate. And Section 5 concludes.

2. Previous results

Here we will put forward some previous propositions and results [1, 2] that are of interest in this paper.

Consider the sum of n stochastic variables x_i

$$S_n = \sum_{i=1}^{n} x_i$$

together with the generalized moments of x_i and S_n, i.e.,

$$\mu_{iq} = \langle x_i^{\,q} \rangle, \forall i \in n, q \in \mathbb{N}$$

and

$$\nu_{nq} = \langle S_n^{\,q} \rangle, \forall n, q \in \mathbb{N}$$

The x_i's are assumed to be identically distributed. Given the probability density function (PDF) of a x_i, i.e., $f_i(x_i)$, it follows from the classic central limit theorem (CLT) that the PDF of its reduced variable will be Gaussian as $n \to \infty$. Necessary conditions for the CLT to hold are as follows. (1) The x_i's are independent, (2) the x_i's have finite second moments, and (3) the infinitesimality condition holds. The latter condition states that

$$\lim_{n\to\infty} \frac{\max\left(\mu_{i2}, i = 1,...,n\right)}{v_{n2}} = 0$$

where $\max\left(\mu_{i2}, i = 1,...,n\right)$ is the largest value of the second moment [3, 4].

A number of authors have tackled the problem of the reduced variable's convergence speed toward its asymptotic Gaussian (e.g., Chebyshev [5], Gnedenko and Kolmogorov [6], Berry [7], and Esseen [8]). One celebrated theorem by Berry and Esseen puts that, under proper conditions, the convergence speed is governed by the absolute value of x_i's third moment over the cube of its standard deviation.

More recently some have employed tools of statistical physics to crunch data coming from subjects as diverse as economics and finance [9, 10] and biology [11]. One work of particular interest is that of Mantegna and Stanley [9]. They put forward a truncated Lévy flight (TLF). The TLF is able to explain several properties observed in economic time series, such as scaling power laws in second moments and slow convergence speed to the Gaussian regime. These are at odds with conventional wisdom but still consistent with the CLT.

We have shown [2] that major particular features of the TLF can be explained in terms of autocorrelations. To see how, we first define an extended nonlinear autocorrelation as

$$\langle q_1 q_2 ... q_k \rangle_n = \sum_{i_1 \cdots i_k = 1}^{n} \left(\langle x_{i_i}^{q_1} ... x_{i_k}^{q_k} \rangle - \langle x_{i_i}^{q_1} \rangle ... \langle x_{i_k}^{q_k} \rangle \right)$$

The usual linear correlation term is obtained from

$$\langle 11\rangle_n = \sum\left(\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle\right) \tag{1}$$

Here there is a power law in the second moment of type $\nu_{n2} \propto n^{1/\alpha}, \alpha \neq 2$, within a finite time window $n_1 \leq n \leq n_2$, even if (1) stands at its "noise" level [2].

In particular, the slow convergence might be caused by nonlinear autocorrelations. And the actual distance of a given distribution $f(x)$ from its ultimate Gaussian state can be measured, as a result. A distance function is defined as [1]

$$D(f, Gauss) = \int_{-\delta}^{\delta} \sqrt{\left|w^f(z)\right|^2}\, dz$$
$$w^f(z) = w_R^f(z) + I w_I^f(z), I = \sqrt{-1}$$

where

$$\begin{cases} w_R^f(z) = -\dfrac{1}{12}(K_n)z^2 + O(z^4), \\ w_I^f(\mathrm{z}) = \dfrac{1}{3}(Sk_n)z + \dfrac{1}{6}(Sk_n)z^3 + O(z^5) \end{cases}$$

Symbols Sk_n and K_n stand for skewness and kurtosis respectively. They are defined as

$$Sk = \left\langle \left(\frac{x - \langle x\rangle}{\mu_2} \right)^3 \right\rangle = \frac{\langle x^3\rangle - 3\langle x^2\rangle\langle x\rangle - 2\langle x^3\rangle}{\left(\langle x^2\rangle - \langle x\rangle^2\right)^{3/2}} \tag{2}$$

and

$$K = \left\langle \left(\frac{x - \langle x\rangle}{\mu_2} \right)^4 \right\rangle - 3 = \frac{\langle x^4\rangle - 4\langle x^3\rangle\langle x\rangle + 6\langle x^2\rangle\langle x\rangle^2 - 3\langle x^4\rangle}{\left(\langle x^2\rangle - \langle x\rangle^2\right)^2} - 3 \tag{3}$$

As for the corresponding sum variable, we have shown that

$$\frac{1}{3}Sk_n = \sum_{i=1}^{n}\frac{\mu_{i2}^{3/2}\mu_{i3}}{3\nu_{n2}^{3/2}} + \frac{\langle 111\rangle_n + 3\langle 12\rangle_n}{\nu_{n2}^{3/2}} \equiv \frac{1}{3}Sk_n^0 + Sk_n^1$$

and

$$\frac{1}{12}K_n = \sum_{i=1}^{n}\frac{\mu_{i2}^{2}\mu_{i4}}{12\nu_{n2}^{2}} + \frac{1}{4}\left(1-\frac{\sigma_{n2}^2}{\nu_{n2}^2}\right) -$$
$$\frac{1}{12}\frac{\langle 1111\rangle_n + 6\langle 112\rangle_n + 4\langle 13\rangle_n + 3\langle 22\rangle_n}{\nu_{n2}^2} \equiv \frac{1}{12}K_n^0 + K_n^1 + K_n^2$$

where

$$\sigma_{nq} = \mu_{1q} + \mu_{2q} + \ldots + \mu_{nq} = \langle x_1^q\rangle + \ldots + \langle x_n^q\rangle$$

Term Sk_n^0 is the skewness of S_n in an independent and identically distributed (IID) process. Term Sk_n^1 involves nonlinear autocorrelations of third order. Even if the correlation of pairs $\langle 11\rangle$ is nil, Sk_n^1 may not be so. In such a case $w_I^f \neq 0$, and this prevents the asymptotic Gaussian regime to be reached. Term K_n^0 is that of an IID process. And term K_n^1 contains linear correlation of pairs. Yet autocorrelations of fourth order appear in K_n^2. Both linear and nonlinear autocorrelations are critical for the understanding of w_R^f. Although linear autocorrelations play a key role in the convergence of a distribution, it is still necessary to take nonlinear autocorrelations into account to fully characterize a process.

Thus the main properties of the TLF can be grasped by an analysis of nonlinear autocorrelations. Such an approach is universal in that it encompasses any stochastic process of finite variance, not only those related to the TLF.

Next section will move on to consider IID *reduced* variables.

3. Reduced variables that are independent and identically distributed

First consider the mean of x_i, $\mu_{i1} = \langle x_i \rangle$, its variance, $\mu_{i2} =< x_i^2 > - < x_i >^2$, and standard deviation $\sqrt{\mu_{i2}}$. The PDF of x_i, say $f_i(x_i)$, is assumed to be distinct for every x_i. Here we are particularly interested in reduced variables, i.e., $\bar{x}_i = \frac{x_i - \mu_{i1}}{\sqrt{\mu_{i2}}}$. Then we define a class of reduced variables that are independent and identically distributed (IIDR) as follows.

Definition 1. Distributions $f_i(x_i)$ of x_i are such that $f_i(x_i) \neq f_j(x_j), i \neq j$. But $\bar{f}_i(\bar{x}_i) = \bar{f}_j(\bar{x}_j)$ for the distributions of any pair of reduced variables $(x_i - \mu_{i1})\big/\sqrt{\mu_{i2}}$ and $(x_j - \mu_{j1})\big/\sqrt{\mu_{j2}}$, $i \neq j$.

Nonidentity for n of these variables is entirely determined by both their means $\mu_{i1}, i = 1, \cdots, n$ and standard deviations $\sqrt{\mu_{i2}}, i = 1, \cdots, n$. A reduced random generator (RRG) G_r is one with zero mean and unit standard deviation. As a result,

$$x_i = \sqrt{\mu_{i2}} G_r + \mu_{i1} \tag{4}$$

An IIDR time series can thus be obtained as follows. (1) Choosing a particular RRG, (2) choosing actual values $\mu_{i1}, i = 1, ..., n$ to capture the mean's time evolution, and (3) choosing actual values $\sqrt{\mu_{i2}}, i = 1, ..., n$ to track the standard deviation's behavior over time.

The results in the previous section can be applied to a time series obtained from an RRG. For instance, we can pick an RRG derived from a TLF of $\alpha = 1$ (Cauchy distribution)

$$f(x)=\begin{cases}\dfrac{C}{1+x^2} & -L_1\le x\le L_2\\ 0 & x<-L_1 \quad\text{or}\quad x>L_1\end{cases},C=\frac{1}{\arctan\left(L_1\right)+\arctan\left(\mathrm{L}_2\right)} \tag{5}$$

where $L_1, L_2 > 0$. The statistical moments of $f(x)$ are easily obtained.

Defining a random generator associated with x, whose distribution function is the $f(x)$ in Eq. (5) is a well-known problem. We can relate x with, say y, which is uniformly distributed within interval $[0,1]$, and use probability conservation to show that

$$x=\tan(\left[\arctan(L_1)+\arctan(L_2)\right]y-\arctan(L_1)) \tag{6}$$

Because y is uniformly distributed in $[0,1]$, x will be distributed in $[-L_1, L_2]$ with a TLF of $\alpha=1$. Reduced variable $\overline{x}=\dfrac{x-\mu_1}{\sqrt{\mu_2}}$ will then be distributed according to a reduced TLF. Finally we define a TLF−RRG of $\alpha=1$ from

$$G_r=\frac{\tan\left(\left[\arctan\left(L_1\right)+\arctan\left(L_2\right)\right]\mathrm{rand}()-\arctan\left(L_1\right)\right)-\mu_1}{\sqrt{\mu_2}} \tag{7}$$

where rand() is a uniform random generator in [0,1]. Generator (7) is a reduced TLF entirely determined by L_1 and L_2. If $L_1=L_2=L$ then the TLF is symmetric. Thus one can define an RRG relying merely on μ_{i1} and $\sqrt{\mu_{i2}}$. Thus, the means and standard deviations of an IIDR process are the only parameters that change during a random generation process.

4. Illustration

Now we will illustrate the above technique with real world data. We take a time series from the daily changes of the British pound−US dollar rate from 5 January 1971 to 4 May 2005 (8615 data points). The heart of our technique is as follows. We divide such a sequence into equal, non-overlapped time periods. Then we compute the means and standard deviations of these periods. For instance, defining a *p*-sized period as a sequence of *p* days is meant that the series of 8615 days will have n_p periods of *p* days that are consecutive and non-overlapped ($n_p \times p = 8615$). We then calculate (for each of these periods) the means and standard deviations using the pound-dollar series. The time evolution of these means and standard deviations are shown in Fig. 1 for *p* = 5 and *p* = 20.

Once p and n_p are defined, we are ready to define the following IIDR random generator:

$$\sqrt{(\mu_{i2})}G_r + A\mu_{i1}, i = 1,2,3,...,8615 \qquad (8)$$

where A is a real number in interval $[0,1]$. The μ_{i1} and $\sqrt{\mu_{i2}}$ are given by

$$\begin{aligned}
&\mu_{11} = \mu_{21} = \mu_{31} = ... = \mu_{p1} = \text{first-period mean} \\
&\mu_{p+11} = \mu_{p+21} = \mu_{p+31} = ... = \mu_{2p1} = \text{second-period mean} \qquad (9) \\
&\mu_{2p+11} = \mu_{2p+21} = \mu_{2p+31} = ... = \mu_{3p1} = \text{third-period mean} \\
&\vdots
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{\mu_{12}} = \sqrt{\mu_{22}} = \sqrt{\mu_{32}} = ... = \sqrt{\mu_{p2}} = \text{first-period standard deviation} \\
&\sqrt{\mu_{p+12}} = \sqrt{\mu_{p+22}} = \sqrt{\mu_{p+32}} = ... = \sqrt{\mu_{2p2}} = \text{second-period standard deviation} \qquad (10) \\
&\sqrt{\mu_{2p+12}} = \sqrt{\mu_{2p+22}} = \sqrt{\mu_{2p+32}} = ... = \sqrt{\mu_{3p2}} = \text{third-period standard deviation} \\
&\vdots
\end{aligned}$$

What generator (8) does is to produce 8615 values as follows. As we generate values for each *p*, the values of μ_{i1} and $\sqrt{\mu_{i2}}$ (that define the IIDR process) alter. This

IIDR generator preserves the time evolution of means and standard deviations related to the *p*-sized periods of the pound-dollar series. This works as if we were reproducing in the random generator the same actual time evolution of means and standard deviations.

Our task is then to compare the statistical properties of the pound-dollar time series with those of an RRG obtained with μ_{i1} and $\sqrt{\mu_{i2}}$ (as defined above). The time evolution of the moments for the two series will be shown to behave similarly, as long as one makes proper choices of A, L_1, and L_2. Accordingly an identically distributed process with autocorrelations can be obtained from an independent, yet nonidentically distributed, random generator.

We employ Eqs. (2) and (3) to capture the time evolution of skewness and kurtosis in the pound-dollar returns. And we take an RRG process that is independent but nonidentically distributed. (We also assume that taking identically distributed variables is harmless.)

Both symmetric ($L_1 = L_2$) and asymmetric ($L_1 \neq L_2$) cases are considered. For robustness, the routine in Eq. (8) is repeated twenty times. In every case we pick a different seed for the uniformly distributed generator in Eq. (7). Figures 2−6 show mean values of 20 processes generated from an IIDR process (as in Eq. (8)). Outcomes for processes with $p = 5$, $n_p = 1723$ (trading weeks of 5 days) and $p = 20$, $n_p = 430$ (trading months of 20 days) are displayed.

Fig. 2−6 present values of standard deviation, skewness, and kurtosis. For completeness, outcomes for a random aggregation of (IID) variables are also shown. And "experimental" outcomes for the pound-dollar returns are shown for comparison.

Fig. 2 presents a symmetric RRG. We get $L_1 = L_2$ from maximum likelihood estimates. Note that kurtosis behavior in the IIDR process is very similar to the actual value. This suggests that kurtosis behavior can be explained in terms of the time evolution of the standard deviation defining the RRG. Skewness behavior is not that clear-cut, however. Yet this is expected because the generator is symmetric and $A = 0$. Note, too, that the standard deviation behaves as if the process had a Hurst exponent of ½.

Fig. 3 shows an asymmetric RRG. Note that the two curves are very similar (as those in Fig. 1). Thus we conclude that kurtosis behavior can be explained by the evolution of the standard deviations. Also, the second moment of the IIDR process (Eq. (17)) is similar to that of an IID process. This suggests lack of linear autocorrelation, despite the fact that nonlinear autocorrelations are surely present. These can be seen from kurtosis behavior.

Fig. 4 displays a symmetric RRG with $A = 1$. $L_1 = L_2$ is chosen as before. Kurtosis of the IIDR process is very similar to that of the actual data. Accordingly a particular time evolution of mean and standard deviation suffices to track the bulk of higher-order moment behavior. Yet the standard deviation cannot fit either the IID or actual data, because its Hurst exponent is lower than that of the pound-dollar series. This particular IIDR process thus presents slow convergence to the Gaussian together with a Hurst greater than ½. These are typical features of the TLF.

Fig. 5 is equivalent to Fig. 4, apart from the fact that it shows a symmetric RRG. Thus the conclusions drawn from Fig. 4 extend to Fig. 5.

Fig. 6 departs from Fig. 5 in that A is set to 0.55. The Hurst exponent is sensitive to changes in A. The Hurst is ½ for $A = 0$, but grows as A increases (not shown). At $A = 0.55$ there occurs the best fit for the standard deviation and kurtosis. This reinforces the standpoint that the generated process may be independent, though nonidentically, distributed.

5. Conclusion

This paper suggests that the main features of nonlinear autocorrelations can be explained in terms of reduced variables that are independent and identically distributed (IIDR). It seems that all relevant information concerning the correlations are encompassed by the time evolution of mean and standard deviation. This makes it possible for a process to be in fact independent though nonidentically distributed. Nonidentity can satisfactorily explain the slow convergence to the Gaussian regime as well as the emergence of a Hurst exponent greater than ½. And it is still possible to observe a non-IID behavior in skewness and kurtosis even if the Hurst equals ½.

Nonconvergence to the Gaussian can thus be explained by departures from the infinitesimality hypothesis of IIDR processes. Second moment is indeed highly volatile.

Thus one should expect the ratio of the highest volatility of each variable and volatility of the cumulative variable to approach zero very slowly, thereby preventing the Gaussian regime to be reached.

From a physicist's perspective, mean and volatility are barometers for market mood. If a market does not change its mood frequently, the infinitesimality hypothesis is likely to hold regardless of how eerie or troubled a market currently stands. Yet a market presenting strong swings in mood impacts volatility very heavily, thereby slowing down convergence to the Gaussian.

All these novel results are in line with our previous findings [1]. Mood swings are more usual in currencies of emerging markets. Yet relative percentage changes in weekly volatilities are less sharp in these currencies. And developed currency markets are less volatile. As a result, sharp swings in volatility causes the breakdown of the infinitesimality hypothesis. And this explains why exchange rates of emerging countries are both more volatile and farer from the Gaussian if compared to those of developed countries. However, sluggishness is stronger in developed countries. This is because a tiny change in a near-zero volatility pushes the limit of the ratio of the highest volatility of each variable and volatility of the cumulative variable toward zero more slowly.

Acknowledgements

The authors acknowledge financial support from the National Council for Scientific and Technological Development (CNPq). I. Gleria also acknowledges support from the Fund of Assistance to Research of Alagoas (FAPEAL).

References

[1] A. Figueiredo, I. Gleria, R. Matsushita, S. Da Silva, On the origins of truncated Lévy flights, Phys. Lett. A 315 (2003) 51-60.

[2] A. Figueiredo, I. Gleria, R. Matsushita, S. Da Silva, Lévy Flights, autocorrelation, and slow convergence, Physica A 337 (2004) 369-383.

[3] V. M. Zolotarev, Reflections on the classical theory of limit theorems, Theory Probab. Appl., 36 (1) (1990), 124-137.

[4] D. Dugué, Oeuvres de Paul Lévy, Vol. III: Eléments Aléatoires, Gauthiers-Villars, Paris, 1976.

[5] P. L. Chebyshev, Sur deux theorems relatifs aux probabilités, Acta Math. 14 (1890), 305-315.

[6] B. V. Gnedenko, A. N. Kolmogorov, Limit distributions for sums of independent random variables, Addison Wesley, Cambridge MA, 1954.

[7] A. C. Berry, The accuracy of the Gaussian approximation to the sum of independent variates, Trans. Amer. Math. Soc. 49 (1941) 122-136.

[8] C. G. Esseen, Fourier analysis of distributions functions. A mathematical study of the Laplace-Gaussian law, Acta Math. 77 (1945) 1-125.

[9] R. N. Mantegna, H. E. Stanley, Stochastic process with ultraslow convergence to a Gaussian: the truncated Lévy flight, Phys. Rev. Lett. 73 (1994) 2946-2949.

[10] R. N. Mantegna, H. E. Stanley, Scaling behavior in the dynamics of an economic index, Nature 376 (1995) 46-49.

[11] G. M. Viswanathan, V. Afanasiev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, H. E. Stanley, Lévy flight search patterns of wandering albatrosses, Nature 381 (1996) 413 – 415.

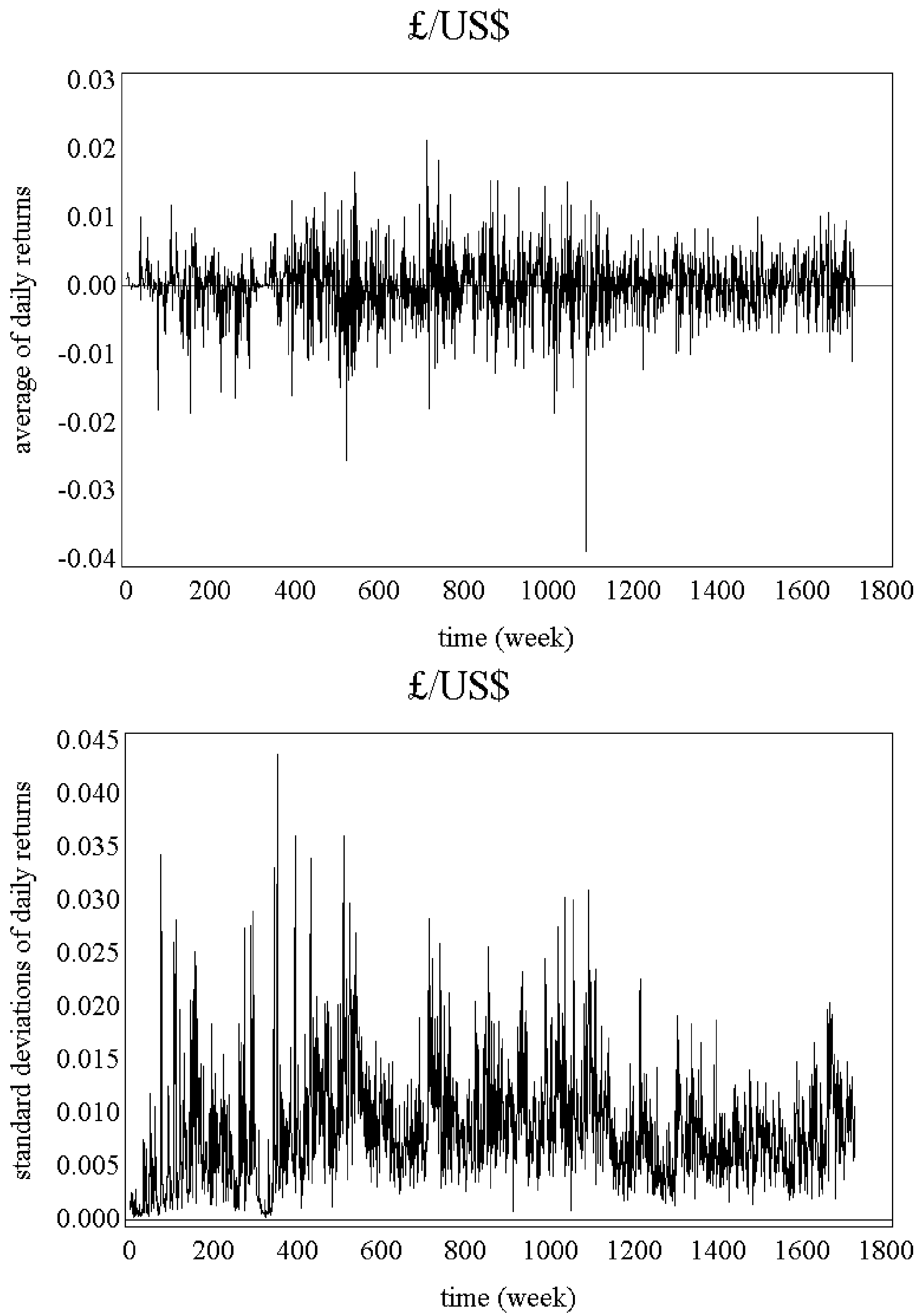

Fig. 1. (a) Time evolution of reckoned means (upper panel) and standard deviations (lower panel) of the daily pound-dollar rate for 1723 trading weeks from 5 January 1971 to 4 May 2005 ($p = 5$).

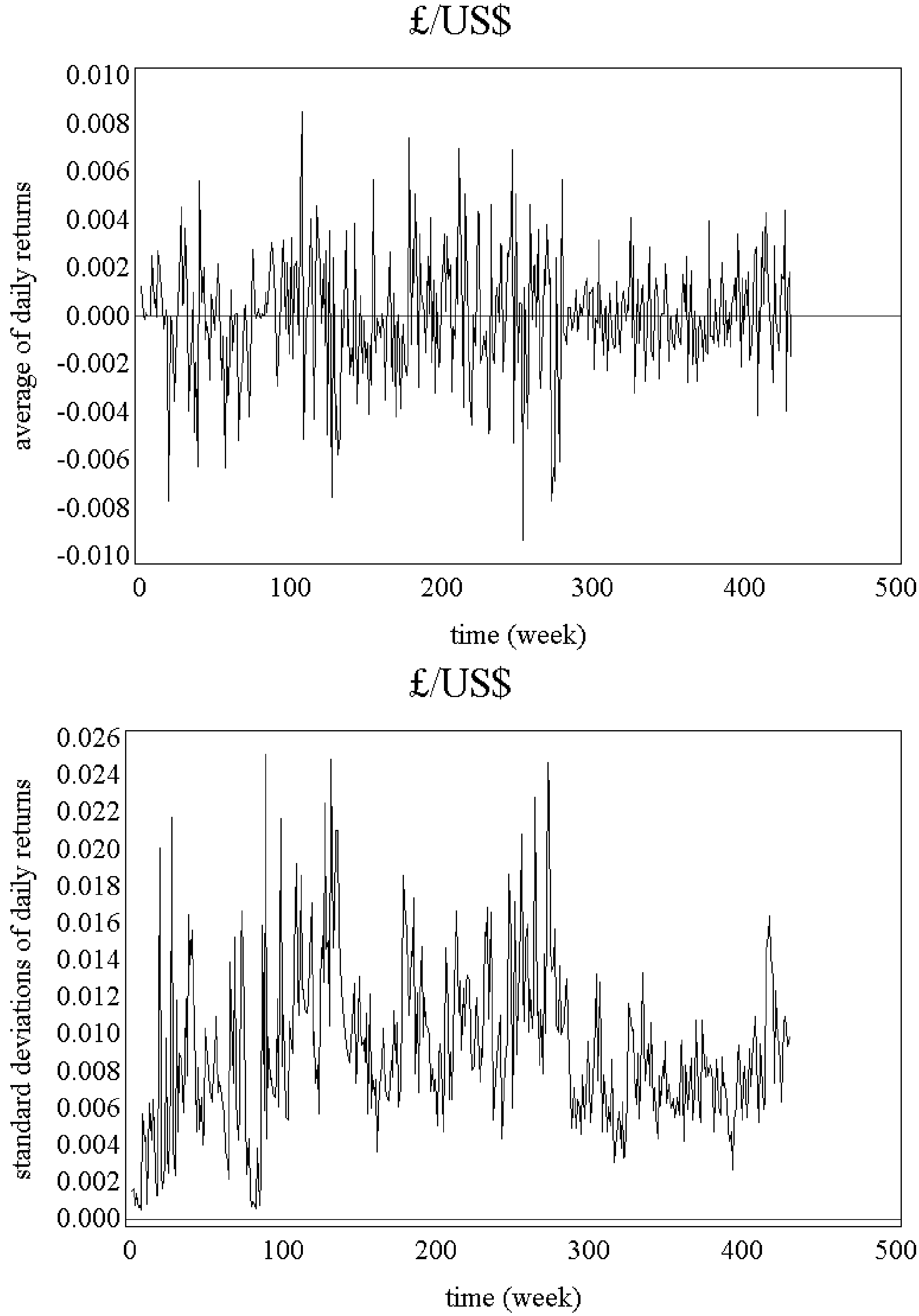

Fig. 1 (b) Time evolution of reckoned means (upper panel) and standard deviations (lower panel) of the daily pound-dollar rate for 1723 trading weeks from 5 January 1971 to 4 May 2005 ($p = 20$).

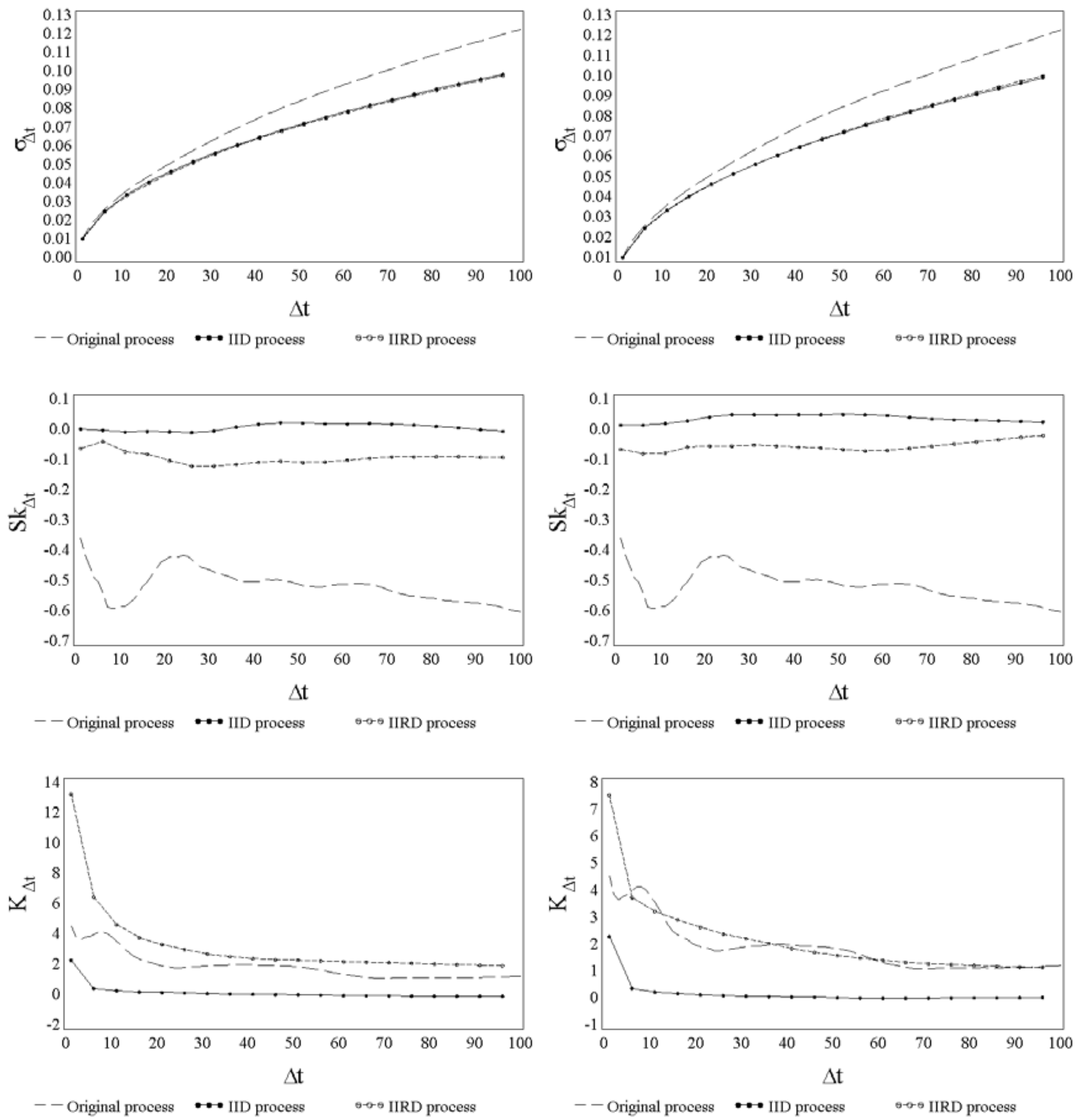

Fig. 2. Scaling in standard deviations (upper panels), skewness (middle panels), and kurtosis (lower panels) of the daily pound-dollar rate. Plots on the left show the IIRD process obtained with p = 5, A = 0, and the symmetric case $L = L_1 = L_2$. Maximum likelihood estimate of L is 7.5. Plots on the right hand side show the IIRD process with p = 20.

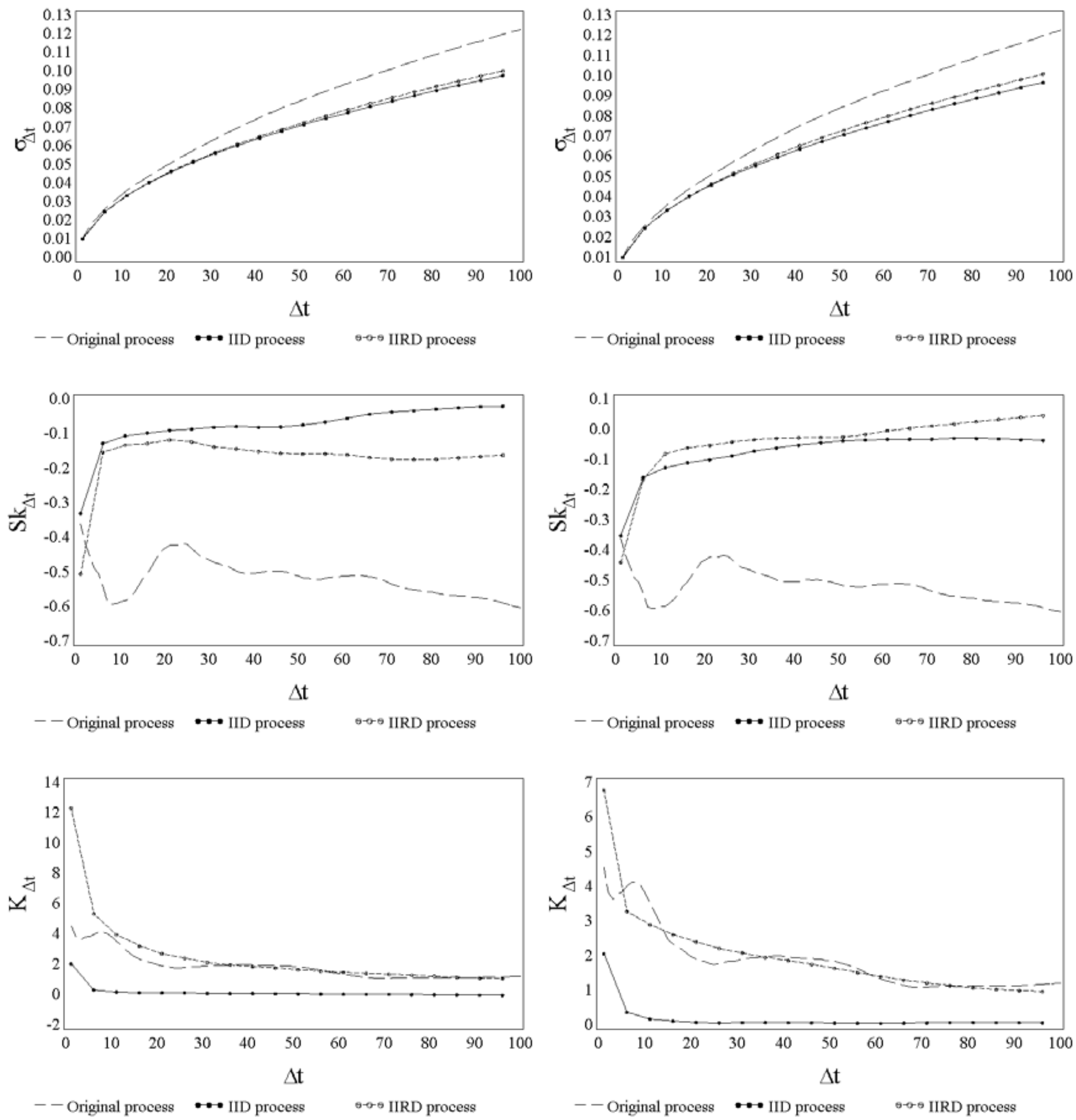

Fig. 3. Scaling in standard deviation (upper panels), skewness (middle panels), and kurtosis (lower panels) of the daily pound-dollar rate. Plots on the left show the IIRD process obtained with p = 5, A = 0, and the asymmetric case $L_1 \neq L_2$. Maximum likelihood estimates of L_1 and L_2 are 7.5 and 6.15 respectively. Plots on the right hand side show the IIRD process with p = 20.

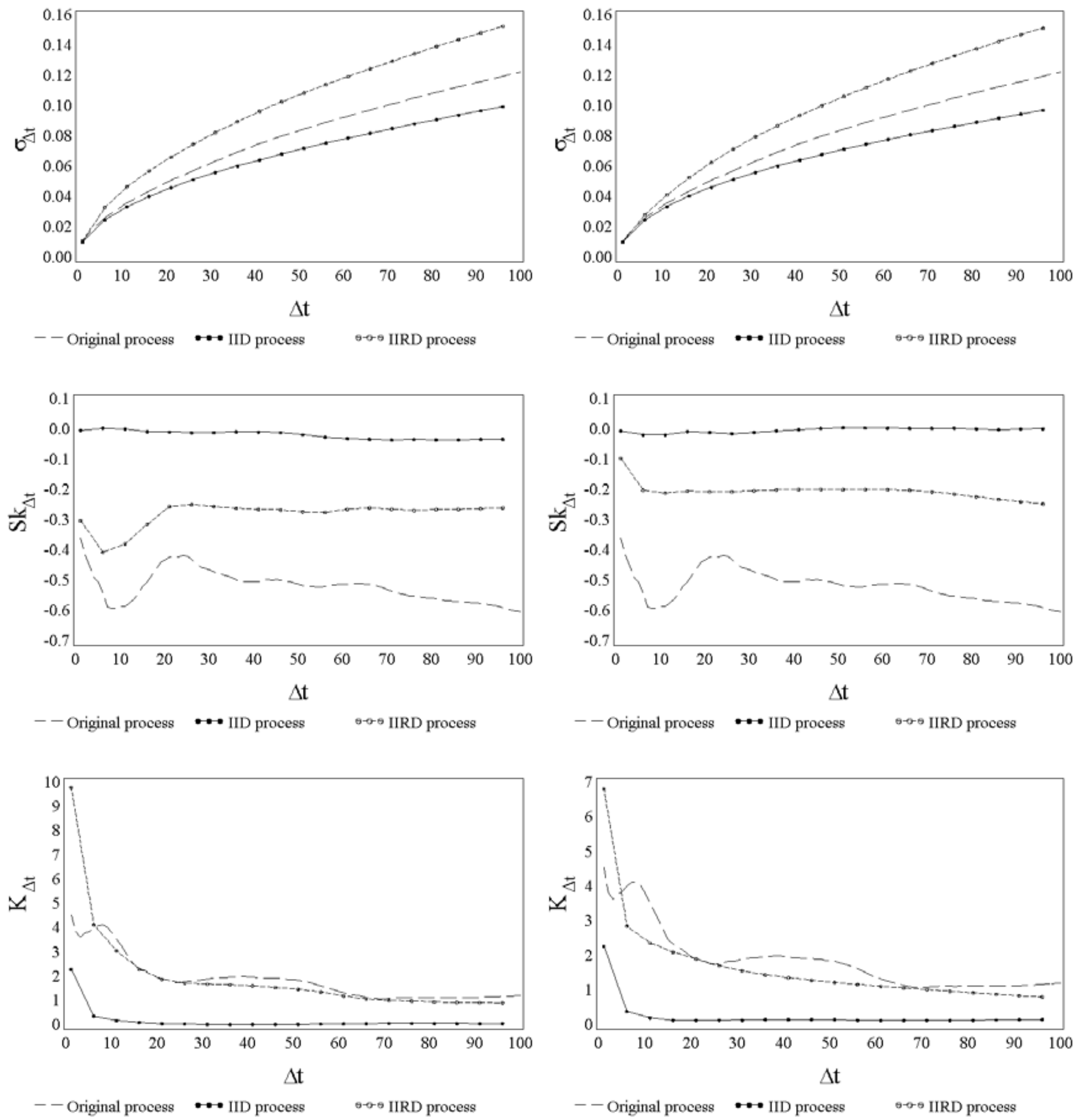

Fig. 4. Scaling in standard deviations (upper panels), skewness (middle panels), and kurtosis (lower panel) of the daily pound-dollar rate. Plots on the left show the IIRD process obtained with p = 5, A = 1, and the symmetric case $L = L_1 = L_2$. Maximum likelihood estimate of L is 7.5. Plots on the right hand side show the IIRD process with p = 20.

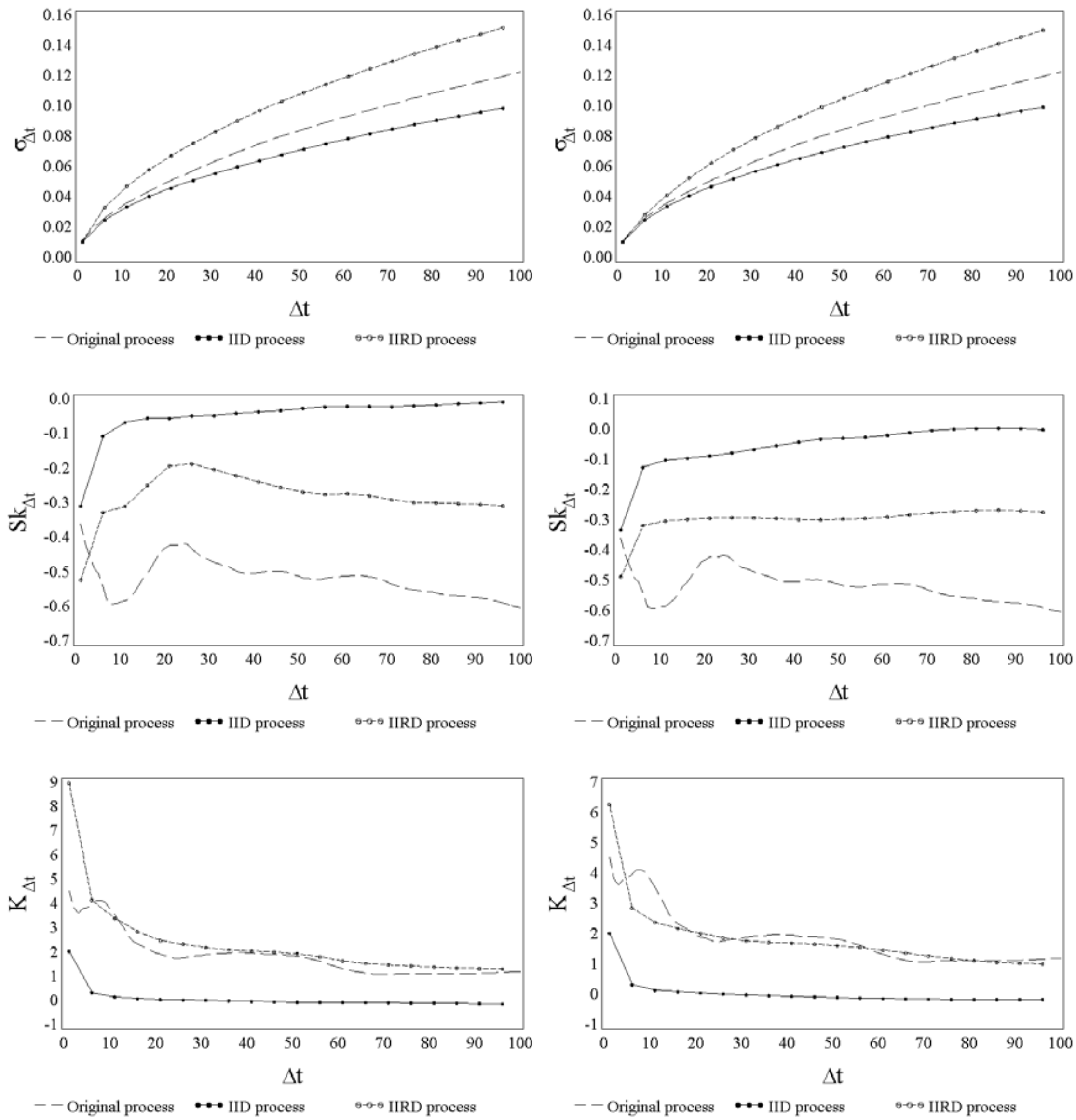

Fig. 5. Scaling in standard deviation (upper panels), skewness (middle panels), and kurtosis (lower panels) of the daily pound-dollar rate. Plots on the left show the IIRD process obtained with p = 5, A = 0, and asymmetric case $L_1 \neq L_2$. Maximum likelihood estimates of L_1 and L_2 are 7.5 and 6.15 respectively. Plots on the right hand side show the IIRD process with p = 20.

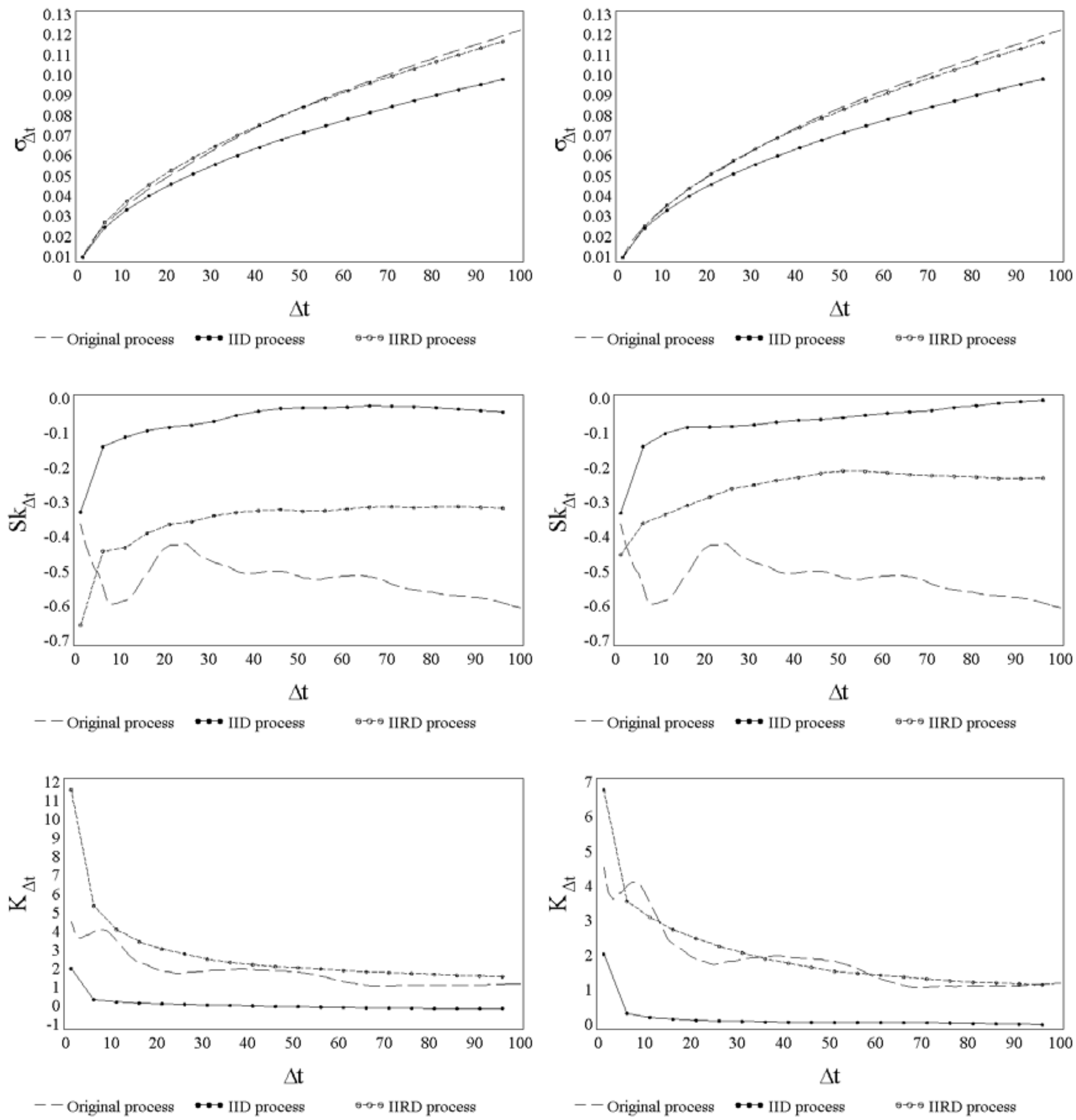

Fig. 6. Scaling in standard deviation (upper panels), skewness (middle panels), and kurtosis (lower panels) of the daily pound-dollar rate. Plots on the left show the IIRD process obtained with $p = 5$, $A = 0.55$, and asymmetric case $L_1 \neq L_2$. Maximum likelihood estimates of L_1 and L_2 are 7.5 and 6.15 respectively. Plots on the right hand side show the IIRD process with $p = 20$.

To be published by *Physica A*