On the Study of Two Models for Integer-Valued High-Frequency Data

Andrea Cremaschi and Jim E. Griffin

Abstract Financial prices are usually modelled as continuous, often involving geometric Brownian motion with drift, leverage and possibly jump components. An alternative modelling approach allows financial observations to take integer values that are multiples of a fixed quantity, the ticksize - the monetary value associated with a single change during the price evolution. In the case of high-frequency data, the sample exhibits diverse trading operations in a few seconds. In this context, the observables are assumed to be conditionally independent and identically distributed from either of two flexible likelihoods: the Skellam distribution - defined as the difference between two independent Poisson distributions - or a mixture of Geometric distributions. Posterior inference is obtained via adaptive Gibbs sampling algorithms. Comparisons of the models applied to high-frequency financial data is provided.

Keywords Time series · High-frequency data · Integer-valued random variables · Bayesian Econometrics · Adaptive MCMC

1 Introduction

The last decades of financial activities have seen rapid technological change in the execution of financial transactions. This has allowed the recent introduction of high-frequency trading (HFT), an automated way of performing transactions based on the use of algorithms interacting with electronic price books, called limit order books (see [5] for an overview of market dynamics and existing statistical literature). The electronic system generating the financial data is characterised by a very high transaction speed, conveying a large volume of data, making high-frequency trading an increasingly promising field for statistical analysis. Indeed, being able to observe the

A. Cremaschi (✉) · J.E. Griffin
Department of Biostatistics,
Universitetet i Oslo, Sognsvannsveien 9, Domus Medica, 0372 Oslo, Norway
e-mail: andrea.cremaschi@biotek.uio.no

J.E. Griffin
e-mail: J.E.Griffin-28@kent.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics in Action*, Springer Proceedings in Mathematics & Statistics 194, DOI 10.1007/978-3-319-54084-9_3

price processes so often in time could provide more information about their dynamics. From a structural point of view, bid and ask prices are characterised by being integer multiples of a fixed quantity, the ticksize, the minimum monetary value tradable in the market. This discrete interpretation of the price processes represents one of the peculiarity of high-frequency data modelling (we refer to [7] for a statistical analysis based on ticksize). In particular, the focus of this work will be on the New York Stock Exchange (NYSE), where the ticksize is of one dollar cent (0.01). Electronic markets such as the NYSE are characterised by their high-depths, meaning that the different bid/ask prices present in the limit order book take values in a wide range (i.e., the *depth* of a market is the number of different prices available for transaction), allowing for market and limit orders to be fulfilled very quickly and hence increase the liquidity while reducing the bid-ask spread (usually equal to a single tick). This feature of the market affects the dynamic of the price processes, influencing the relationships between returns at different time points. This paper is structured as follows: the next Section will introduce the object of the study, the integer-valued returns, and two conditional distributions used to model them. Section 3 will introduce the algorithms used to perform posterior inference, while an application to real stock market data is presented in Sect. 4. Section 5 concludes.

2 Distributions for Tick Data

The nature of the prices makes them interpretable as positive integers, indicating the multiple of the ticksize at which a specific transaction takes place. Let p_t be the price process at time $t > 0$, and P_t the discretised version of it, simply computed as $P_t = \frac{p_t}{ticksize}$. It is assumed that $T > 0$ transactions are observed in a finite time interval, hence the observed prices are $P_1, \ldots, P_T$, taking values over the set of positive integers $\mathbb{N}$. Despite the notation, time dependency will not be assumed in the analysis, in order to present a preliminary study of the alternative discrete transformation of the data, together with an outline of the pros and cons related with the different modelling choices. It is of interest to study the behaviour of the price process, with particular attention to its fluctuations in terms of variability. In order to do so, in Financial Econometrics the concept of a return is introduced: a transformation of the raw price data that is able to provide information about the variability of the data generating process underlying the observations. When the data are assumed to be continuous, the standard definition of return is the log-ratio of two consecutively observed prices. Clearly, this is not useful in the context of this work, due to the discretisation adopted at the basis of the study. Hence, following [3], the ticksize returns are defined as the difference of two consecutive discrete valued prices, such that for each $t > 0$, $Y_t := P_t - P_{t-1} \in \mathbb{Z}$. While, in the continuous case, the returns can be easily modelled by using a Normal distribution, this alternative definition of returns requires a suitable probability distribution defined on $\mathbb{Z}$ in order to build a valid statistical model. In this work, two such distributions are specified, namely the Skellam distribution and the Folded Geometric distribution, and their

inference properties are compared. The first distribution is defined as the difference of two independent Poisson distributed random variables, which has been used, among others, in the analysis of intra-day high-frequency trading data in the work of [4, 10]. A feature of their model is the presence of a zero-inflation parameter to accommodate a mode at zero. This choice is motivated by the high-depth of the electronic market. In this work, a similar issue is tackled by the introduction of the Folded Geometric distribution, defined as a mixture of two Geometric distributions on the positive and negative integers, with two additional parameters to represent centrality and modality. In the rest of this section, the two distributions are introduced, as well as the two models arising from these choices.

2.1 Skellam Distribution

With reference to [3], the integer returns are modelled by using the difference between two positive quantities representing the positive and the negative jumps governing the price evolution, such that $Y_t = L_t^+ - L_t^-$, with $Y_t \in \mathbb{Z}$, for $t = 1, \ldots, T$. Notice that here the jump processes L_t^+ and L_t^- are used to support the modelling choice that will follow, and have a different meaning from the integer-valued prices P_t and P_{t-1} defined above. The two independent distributions can be interpreted as responsible for the evolution of the returns, by making it move up (L_t^+) or down (L_t^-). When $\{L_t^+\}_{t=1}^T$ and $\{L_t^-\}_{t=1}^T$ are two independent Poisson distributions of intensities ϕ^+ and ϕ^-, then Y_t is Skellam distributed, with probability mass function as follows:

$$\mathbb{P}(Y_t = k) = e^{-t(\phi^+ + \phi^-)} \left(\frac{\phi^+}{\phi^-}\right)^{k/2} I_{|k|}(2t\sqrt{\phi^+\phi^-}),$$
$$I_k(x) = \left(\frac{1}{2}x\right)^k \sum_{n=0}^{\infty} \frac{(\frac{1}{4}x^2)^n}{n!(n+k)!},$$

where $I_k(x)$ is the modified Bessel function of the first kind of positive arguments x and k (see [1]). Figure 1a presents the p.m.f.'s of the Skellam distribution for different combinations of the intensity parameters. In this work, we consider an alternative parameterisation of the Skellam distribution in terms of the variance and skewness of the distribution, such as:

$$\begin{array}{ll} \phi^+ = \frac{1+a}{2} e^h \\ \phi^- = \frac{1-a}{2} e^h \end{array} \quad \rightarrow \quad \begin{array}{ll} a = \frac{\phi^+ - \phi^-}{\phi^+ + \phi^-} & = \frac{\mathrm{E}(L)}{\mathrm{Var}(L)} \\ h = \log(\phi^+ + \phi^-) & = \log(\mathrm{Var}(L)) \end{array}$$

Notice that the newly introduced parameters can be interpreted using the moments of the Skellam distribution. In particular, the real-valued parameter h represents the log-volatility of the distribution, while $a \in (-1, 1)$ can be seen as a scaled skewness parameter, since when $Y \sim Sk(\phi^+, \phi^-)$, then $\mathrm{Skew}(Y) = \frac{\phi^+ - \phi^-}{(\phi^+ + \phi^-)^{3/2}} = e^{-h/2} a$.

A computational issue with the Skellam model is the dependence on the modified Bessel function of the first kind, $I_k(x)$. The computation of this hypergeometric series greatly affects the accuracy of the computed probabilities hence influencing the inference. This is usually the case for large values of (k, x), such as when a rapid change in price is observed, i.e. $k = |y_t|$, or when the intensity parameters are such that $x = \sqrt{2\psi^+\psi^-}$ takes large values. It is worth noting that, in the latter case, large values of x correspond to large values of the variance of the returns, but could as well be associated to small values of the conditional mean. In order to avoid this problematic aspect, a latent variable is introduced in the model to represent the negative jumps in the evolution of the returns, $\{L_t^-\}_{t=1}^T$. The resulting likelihood for the returns is therefore a shifted Poisson, shifted by $-L_t^-$ units. In a Gibbs sampler targeting the posterior distribution, this yields a much easier expressions for the full conditionals of the parameters, and hence eases the computation. The final Skellam model analysed in this work is the following:

$$
\begin{aligned}
&Y_t|L_t^-, h, a \overset{\text{ind}}{\sim} ShPoi\left(-L_t^-, \frac{1+a}{2}e^h\right), && h \sim N(\mu, \psi),\\
&L_1^-, \ldots, L_T^-|h, a \overset{\text{iid}}{\sim} Poi\left(\frac{1-a}{2}e^h\right), && \mu \sim N(0, 1),\\
&\frac{a+1}{2} \sim Beta(0.5, 0.5), && \psi \sim inv-gamma(3, 2),
\end{aligned}
$$

where $X \sim ShPois(s, \eta)$ is distributed according to a shifted Poisson with shifting parameter s and intensity η, if $(X - s) \sim Poi(\eta)$. Moreover, $N(m, s^2)$ indicates the normal distribution with mean m and variance s^2, and $inv-gamma(a, b)$ indicates the inverse-Gamma distribution with mean $\frac{b}{a-1}$ and mode $\frac{b-1}{a+b-2}$. Notice how the parameter a is modelled as a linear transformation of the Beta distribution $Beta(a, b)$ with mean $\frac{a}{a+b}$ and mode $\frac{a-1}{a+b-2}$, as done in [9].

2.2 Folded Geometric Distribution

As mentioned before, it is useful to avoid the computation of the Bessel function $_k(x)$. An alternative modelling approach defines a different distribution to model the returns. As mentioned above, the market considered in this work is a one-tick high-depth market (such as the New York Stock Exchange market), meaning that the bid-ask spread is usually equal to one tick, producing transaction returns that fluctuate very little, and that present a lot of zeros. In order to capture this behaviour, a probability distribution F is introduced satisfying the following requirements:

(a) it has support on $\mathbb{Z}$,
(b) it allows the presence of a mode at zero,
(c) it does not include convolution terms (such as the Bessel function $I_k(x)$),
(d) it is flexible enough to represent the evolution of ticksize normalised returns.

To start, define the discrete random variables X^+ and X^- with support on $\mathbb{N} \setminus 0$ and $\mathbb{Z} \setminus \mathbb{N}$, distributed according to F^+ and F^-, respectively. Hence, assume that the two distributions F^+ and F^- admit probability measures indicated with $\mathbb{P}^+$ and $\mathbb{P}^-$. Let X be a discrete random variable defined on $\mathbb{Z}$ with the following p.m.f.:

$$\mathbb{P}(X = k) = \frac{1}{c} \begin{cases} \mathbb{P}^-(X^- = k + l) & k < l \\ a & k = l \\ \mathbb{P}^+(X^+ = k - l) & k > l \end{cases}$$

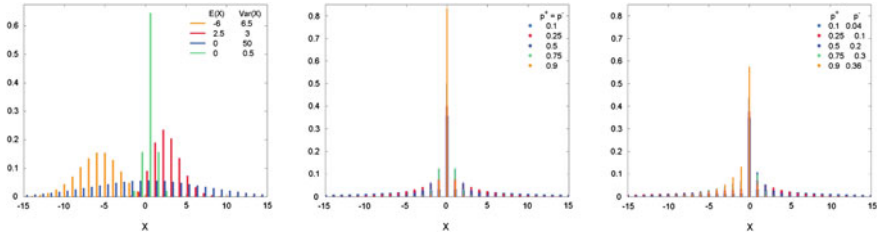
where a is proportional to the probability of taking the value l, representing the centre of the distribution. The term c is the normalising constant of the distribution, and is equal to $c = 2 + a - \mathbb{P}^+(X^+ = l) - \mathbb{P}^-(X^- = l)$. The mixture of these three random variables covers the whole sample space $\mathbb{Z}$, satisfying (a), and can be constructed such that there is a mode at $l = 0$, satisfying (b). Condition (c) and (d) are also satisfied, since this is a mixture density without any convolution, and the two halves can be chosen arbitrarily, providing suitable flexibility for different applications. The resulting distribution is hereby called a *Folded* distribution. In this work, a mixture of two Geometric distributions with success probabilities denoted as p^+ and p^- is considered, together with a mode at $l = 0$, and it will be called *Folded Geometric* distribution, indicated as $FG(p^+, p^-, l, a)$. The first three moments of the random variable X, when $l = 0$, are:

$$\begin{aligned} \mathrm{E}(X) &= \frac{\mathrm{E}(X^+) - \mathrm{E}(X^-)}{c}, \\ \mathrm{E}(X^2) &= \frac{\mathrm{E}(X^{+2}) + \mathrm{E}(X^{-2})}{c}, \\ \mathrm{E}(X^3) &= \frac{\mathrm{E}(X^{+3}) - \mathrm{E}(X^{-3})}{c}. \end{aligned}$$

Consider using this p.m.f. to describe the distribution of the returns. For $t = 1, \ldots, T$:

$$\mathbb{P}(Y_t = k) = \frac{1}{c} \begin{cases} p^-(1 - p^-)^{l-k} & k < l \\ a & k = l \\ p^+(1 - p^+)^{k-l} & k > l \end{cases}$$

notice that $a \geq \frac{1}{4}$ guarantees the unimodality of the distribution at $l = 0$, since $p(1 - p) \leq \frac{1}{4}$ for the Geometric distribution, and that the normalising constant is $c = 2 + a - p^- - p^+$. Finally, notice how the choice of the two success probabilities is completely arbitrary, and no restriction is imposed, apart from the obvious $p^+, p^- \in (0, 1)$. Figure 1b and c show the different shapes of the Folded Geometric distribution, when the Symmetric ($p^+ = p^-$) or Asymmetric ($p^+ \neq p^-$) setting is chosen. Finally, the Folded Geometric model can be outlined as:



(a) *Skellam* (b) $FG(p, 1-p, 0, 1)$ (c) $FG(p, p/2.5, 0, 3.5)$

Fig. 1 Different p.m.f.'s used to describe the conditional distribution of the integer-valued returns

$$y_1, \ldots, y_T | h^+, h^-, a \overset{\text{iid}}{\sim} FG \quad \left(\frac{e^{h^+}}{1+e^{h^+}}, \frac{e^{h^-}}{1+e^{h^-}}, l = 0, a \right),$$

$$\left(a - \tfrac{1}{4}\right) \sim Exp(\tfrac{3}{4}),$$

$$h^+ \sim N(\mu^+, \psi^+), \qquad h^- \sim N(\mu^-, \psi^-),$$

$$\mu^+ \sim N(0, 1), \qquad \mu^- \sim N(0, 1),$$

$$\psi^+ \sim inv - gamma(3, 2), \qquad \psi^- \sim inv - gamma(3, 2).$$

Notice how the centrality parameter a is set to be positive, and greater than the value $\frac{1}{4}$, in order to guarantee the unimodality of the conditional distribution of the observations. Prior specification is analogous to the one used for the Skellam model.

3 Algorithms

Posterior computations for the specified models are obtained via the implementation of adaptive Gibbs sampler algorithms, outlined in this section. The adaptive part of the algorithm scheme is an implementation of the adaptive random walk Metropolis algorithm described in [2]. In their work, the authors present an innovative sampling scheme for the random walk Metropolis-Hastings, where the proposal kernel is allowed to depend on the history of the process, i.e. to depend from the previously sampled values of the parameters. An extensive review of various adaptive algorithms can be found in [8]. To give an illustration of such a procedure, the **AMH** algorithm is reported in the frame below, outlining the sequence of steps necessary to update a non-conjugate parameter q of a given model.

Adaptive random walk Metropolis-Hastings algorithm

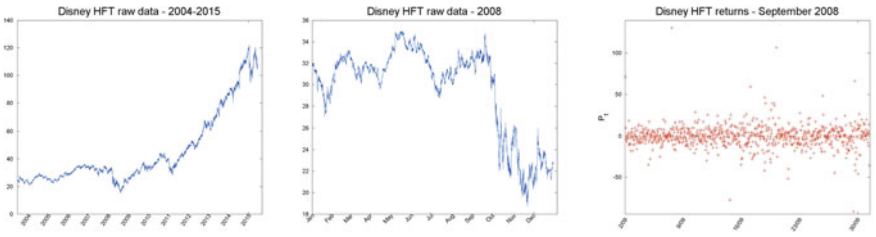
Choose a burn-in value g_0, and initialise θ at iteration $g = 1$;
run $g_0 > 0$ iterations with a fixed value of the proposal variance s_θ^2;
for $g > g_0$, perform the following log-scale update: $\log(s_\theta^2$
$(g + 1)) = \log(s_\theta^2(g)) + (g^{-0.55}(\alpha_\theta - \overline{\tau}))$.

In the algorithm, α_θ is the acceptance rate of the Metropolis-Hastings step, and $\overline{\tau}$ is a reference optimal acceptance rate fixed equal to the value 0.234, following the work of [11, 12]. The parameters for which it is possible to adopt the **AMH** algorithm are the non-conjugate ones, that is a and h in this work (for all the models).

4 Application

In this Section, we present an application to a subset of the Disney transaction data, originally sampled every minute during the years 2004–2015 from the New York Stock Exchange market, and here restricted to the months of September and thinned each 10 min. The reduction and the thinning are adopted in order to reduce the computational burden. Figure 2a shows the whole dataset for the Disney stock (2004–2015), while Fig. 2b shows the year 2008 only, from which the month of September (Fig. 2) is extracted for the analysis.

Gibbs sampler algorithms are run for 525.000 iterations, of which 500.000 constitute the burn-in period, and 5000 are subsequently saved every fifth iteration.The



(a) from 2004 to 2015 (b) 2008 (c)

Fig. 2 Disney HFT raw data

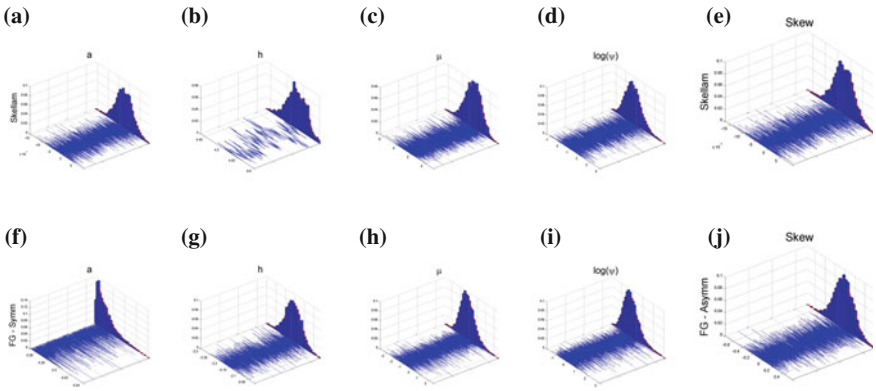
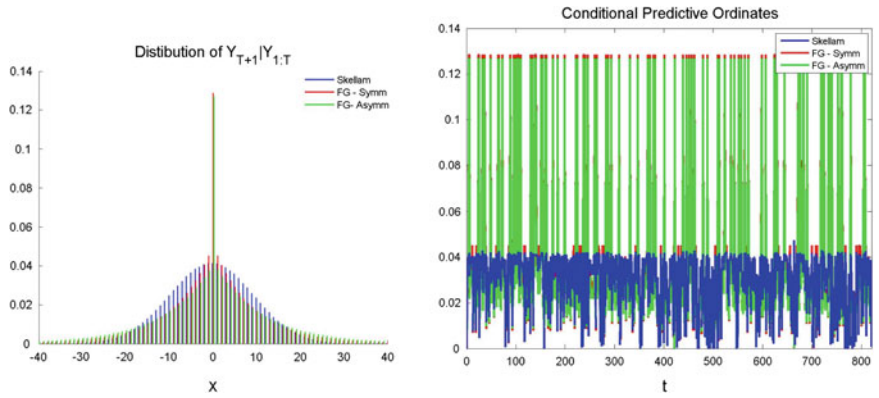


Fig. 3 Traceplots of the posterior MCMC chains for some parameters of the models described. **a–d** Skellam model **f–i** FG-Symmetric model **e** and **j** Skewness of Skellam and FG-Asymmetric model, respectively

Table 1 Acceptance rates and predictive indices log-**BF** and **LPML**

Acceptance rates		log-**BF** \ **LPML**	Skellam	FG - Symm	FG - Asymm
a	0.22602	Skellam		−170.323	−157.1342
h	0.23423				
a	0.23407	FG - Symm	$2.9389 \cdot 10^3$		13.1887
h	0.2352				
a	0.23359	FG - Asymm	$2.9213 \cdot 10^3$	−17.6052	
h^+	0.23356				
h^-	0.2341				



(a) Predictive distribution under different models

(b) CPO under different models

Fig. 4 Predictive comparisons

resulting traceplots with histograms for the Skellam and the Folded Geometric Symmetric models are presented in Fig. 3. Furthermore, a summary of the average acceptance rates is provided in Table 1 (first column). From inspection of the traceplots, it can be argued that the posterior chains have converged for all the parameters, except for the parameter h in the Skellam model, clearly affected by the introduction of the latent variables $L_1^-, \ldots, L_T^-$ (the Asymmetric case is not reported here for unavailability of space, but leads to the same conclusions as the Symmetric case), while the values of the acceptance rates are all very close to the gold standard value 0.234, suggesting that the adaptive algorithms have reached stability. Moreover, density estimation is presented in Fig. 4a, where the predictive distributions for the three different models are displayed. A clear difference in the tails of the distributions can be observed. In particular, the Skellam distribution seems to allocate more mass away from zero, possibly because of the aggregation effect induced by thinning the observations, and the ability of the Skellam distribution to approximate the Normal

distribution when the intensity parameters are equal (see [3] for an analogous result on the Skellam process). Not surprisingly, the Folded Geometric models are instead able to capture the zero-inflated aspect of the distribution of the returns much better than the Skellam model does. Further analysis considered the choice of the most suitable model among the suggested ones, evaluating the *Log Pseudo-Marginal Likelihood* (LPML), as defined by [6] in terms of the *Conditional Predictive Ordinate*:

$$(CPO_t^j)^{-1} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{f^j(y_t|\theta^j(g))}$$

$$LPML_j = \sum_{t=1}^{T}\log(CPO_t^j),$$

where $G = 5000$ is the number of iterations saved, $f^j(y_i|\theta^j(g))$ is the likelihood function for the j-th model, and $\theta^j(g)$ is the g-th MCMC sample of the parameter vector for the j-th model. The higher the values of the LPML, the better the fitting of the data to the j-th model. As it can be seen from Fig. 4b, the Skellam model has higher CPO values for some of the observations, while the two Folded Geometric models look more stable, and in agreement with each other. A measure indicating whether a model is suitable to describe the data at hand is indeed the log-ratio of LPMLs. These values are reported in the right hand side of Table 1, together with the estimates of the log-Bayes Factors for each pair of models. From such values, it appears that the Folded Geometric models are to be preferred to the Skellam model, probably as a consequence of introducing the latent variables $L_1^-, \ldots, L_T^-$. Between the two Folded Geometric models, it seems like the Symmetric one is performing better than the Asymmetric one, suggesting little evidence of asymmetry in the data, as it is also shown by the traceplots of the skewness parameters for the Skellam and Folded Geometric Asymmetric models, in Fig. 3e and j.

5 Discussion

In this work, two different models for the statistical analysis of discretised high-frequency data are presented. The conditional distribution of the observables in the two scenarios is set to be either the Skellam distribution or the Folded Geometric distribution, the latter being in turn distinguishable between its Symmetric and Asymmetric case. An adaptive Gibbs sampling algorithm is described that is able to provide good mixing properties for the posterior chains of the parameters of the different models. Model comparison revealed some discrepancies between the performances of the different models. In particular, the predictive distribution of $Y_{T+1}|\boldsymbol{Y}_{1:T}$ seems to be quite different for the Skellam and the Folded Geometric models, probably due to the heavier-tailed Geometric mixture distribution, that is capable of capturing more extreme behaviours and outliers in the returns, while the predictive distribution for

the Skellam model is closer to a Normal distribution centered at zero. As expected, the Folded Geometric distribution is able to capture the zero-inflated aspect of the returns, differently from the Skellam one. Furthermore, some predictive quantities such as log-LPML and log-Bayes Factor are compared, supporting the idea that the Folded Geometric model might be a better choice for the analysis of high-frequency data when no time dependency is included in the parameter space, this point being crucial in interpreting the results obtained so far. The results provided by the Skellam model can be explained by recalling that the Skellam distribution can be seen as a discretised version of the Normal distribution. On the contrary, in the Folded Geometric case, there is space for detection of extremal behaviours. To conclude, the analysis presented in this work has shown how, under suitable algorithmic conditions and standard prior elicitation choices, the assumption of independent and identically distributed data is accommodated in different ways by different model choices. In order to deepen the study of this matter, it is our intention to study the property of models where time dependence is included at parameter level, as well as in the likelihood term, via the introduction of a stochastic volatility process. In this case, the two models might provide more consistent results, not so distant from one another.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York (1972)
2. Atchadé, Y.F., Rosenthal, J.S.: On adaptive Markov chain Monte Carlo algorithms. Bernoulli **11**(5), 815–828 (2005)
3. Barndorff-Nielsen, O.E., Pollard, D.G., Shephard, N.: Integer-valued Lévy processes and low latency financial econometrics. Quant. Finance **12**(4), 587–605 (2012)
4. Barra, I., Koopman, S.J.: Bayesian dynamic modeling of high-frequency integer price changes. Tinbergen Inst., 16-028 (2016)
5. Cont, R.: Statistical modeling of high-frequency financial data. IEEE Signal Process. Mag. **28**(5), 16–25 (2011)
6. Geisser, S., Eddy, W.F.: A predictive approach to model selection. J. Am. Stat. Assoc. **74**(365), 153–160 (1979)
7. Griffin, J.E., Oomen, R.C.A.: Sampling returns for realized variance calculations: tick time or transaction time? Econ. Rev. **27**(1–3), 230–253 (2008)
8. Griffin, J.E., Stephens, D.A.: Advances in Markov chain Monte Carlo. In: Bayesian Theory and Applications, pp. 104–144 (2013)
9. Kim, S., Shephard, N., Chib, S.: Stochastic volatility: likelihood inference and comparison with ARCH models. Rev. Econ. Stud. **65**(3), 361–393 (1998)
10. Koopman, S.J., Lit, R., Lucas, A.: The dynamic Skellam model with applications (2014)
11. Roberts, G.O., Rosenthal, J.S.: Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**(4), 351–367 (2001)
12. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**(1), 110–120 (1997)