Journal of the Chinese
Statistical Association
Vol. 59, (2021) 98–128

Modelling Over-Dispersion in Price Jumps Arrivals: A Comparison Between Mixture of Poisson and Linear Hawkes Model

Ping-Chen TSAI[†]

Department of Finance, National Sun Yat-sen University

ABSTRACT

Price jumps may display some degree of clustering and this feature is typically manifested through a phenomenon known as over-dispersion in count data. The time series of counts will see a variance larger than its mean, and hence a simple Poisson null hypothesis should be rejected. Two competing models – mixture of Poisson and linear Hawkes models − are shown to be able to reproduce the over-dispersion feature, but the former by definition has a clustering parameter $CP = 0$, whereas the latter has $CP > 0$. Different versions of the two models are fitted to high-frequency price jumps data, and a comparison on estimated degree of over-dispersion is made possible by using a non-parametric CP approximation from linear Hawkes model. A simulation study and a residual analysis are conducted to check the robustness of results.

Key words and phrases: Price jumps, Over-dispersion, Mixture of poisson model, Linear hawkes model.

JEL classification: C13, G10.

[†]Corresponding to: Ping-Chen TSAI
E-mail: vincenttsai@g-mail.nsysu.edu.tw

1. Introduction

The degree of price jumps clustering has been a focus of research in high-frequency financial econometrics. Previous studies using parametric volatility models have suggested a time-varying intensity of jumps in asset prices; see, for example, Fulop, Li and Yu (2015) and Maneesoonthorn, Forbes and Martin (2017). When price jumps display clustering in arrivals, a so-called over-dispersion phenomenon can be observed. Specifically, the time series of counts will see a greater variance than its mean, and thus the null hypothesis of a Poisson model will be rejected[1]. One important consequence of price jumps clustering is on dynamic hedging activities – hedging errors will occur if one assumes a constant price jumps intensity whereas in reality price jumps cluster (Hainaut and Moraux 2018).

The objective of this paper is to compare two methods – mixture of Poisson and linear Hawkes model – for reproducing the over-dispersion phenomenon[2]. In mixture of Poisson model, the intensity of arrivals depends on a state variable and hence is not constant. The linear Hawkes model of Hawkes (1971) has been a popular model in the finance and econometrics literature for the past few years, and it also has a specification of time-varying intensity. Both approaches are able to describe the over-dispersion or clustering of arrivals in count data. However, as pointed out by Filimonov and Sornette (2015), mixture of Poisson model by definition has a clustering parameter $CP = 0$, whereas linear Hawkes model has strictly positive $CP > 0$. If price jumps are modelled as Hawkes process but in fact over-dispersion is generated by mixture of Poisson model, there could be bias or inaccurate model implications for derivative pricing. This study therefore conducts an empirical study to see if it is possible to distinguish the two models in describing over-dispersion in price jumps.

To obtain credible price jump data, a high-frequency jump test is implemented with care to reduce the possibility of detection errors, which are typically encountered in a high-frequency, intraday environment. The price jumps are obtained from a sample period of 2,076 complete trading days, each sampled at 5-minute and 2-minute intervals.

[1]See Cox and Isham (1980), Daley and Vere-Jones (2003) and Zucchini and Macdonald (2009).

[2]Other models that can produce over-dispersion include a negative binomial distribution or a zero-inflated Poisson (ZIP) model; see the review by Giles (2010) and Rodriguez (2013).

The daily intensities of price jumps are 20.04% and 41.23% respectively for the 5-minute and 2-minute frequencies. Consistent with previous literature, the price jumps data displays some degree of clustering, as manifested by the larger-than-one dispersion ratios in daily counts and in the inter-arrival times between identified jumps. A simple Poisson model therefore is not sufficient for describing the price jump arrivals[3].

In search for a satisfactory model for the price jumps, different versions of mixture of Poisson and linear Hawkes models are considered. In the literature, there are many ways in determining the number of components m in a finite mixture model; see the review by Oliveira-Brochado and Martins (2005). For simplicity, mixture of Poisson models with $m \leq 3$ components are estimated in this paper. In light of the many zero observations in the count data of price jumps, a zero-inflated Poisson (ZIP) model is also considered. On the other hand, linear Hawkes model can be specified with an exponential or an inverse-power decay function $\omega(\tau)$, where $\tau \geq 0$ is the elapsed time since previous events. An exponential decay function has a faster rate of decline than its inverse-power counterpart[4]. As pointed out by Hardiman and Bouchaud (2014) and Filimonov and Sornette (2015), the estimated CP value is sensitive to the choice of decay function. We design the two decay functions to have the same number of parameters, and model selection can be made based on maximized log-likelihood value. Ferriani and Zoi (2020) also describe the clustering of detected price jumps using linear Hawkes model; however, they only consider an exponential decay function.

To evaluate the models' ability in describing the over-dispersion phenomenon, for the Hawkes model we use the non-parametric estimator of CP in Hardiman and Bouchaud (2014) and convert estimated CP values into dispersion ratios. For mixture of Poisson model, the dispersion ratio can be calculated directly using the model's expected value and variance. This approach circumvents the problem that mixture of Poisson and linear Hawkes model are non-nested, which renders a direct comparison between them – such as a likelihood ratio test – infeasible.

Our estimation results show that for 5-minute price jumps a ZIP model is pre-

[3]Many parametric volatility models are, however, augmented with Poisson-type price jumps to account for sudden and large price movements; see for example, Eraker, Johannes and Polson (2003).

[4]See the discussion in Bauwens and Hautsch (2009).

ferred, whereas for 2-minute price jumps both ZIP and mixture of Poisson model with $m = 2$ can be considered. In estimating linear Hawkes model, we obtain stable CP values from the two decay functions and calculate the standard errors of CP. For both 5-minute and 2-minute price jumps, the exponential decay function gains higher maximized log-likelihood. Thus for 5-minute price jumps, ZIP is compared with linear Hawkes model with exponential decay in reproducing over-dispersion; both models give values larger than is observed from data, with ZIP closer to the target. For 2-minute price jumps, the comparison is between mixture of Poisson with m=2 and linear Hawkes with exponential decay. The former produces an over-dispersion value that matches the empirical target, whereas the latter is biased upward. Mixture of Poisson model, with ZIP as special case, therefore performs better in reproducing the over-dispersion of price jumps than linear Hawkes model.

Two diagnostic checks are provided for the above conclusion. The first is a residual analysis on linear Hawkes model developed by Ogata (1988), which shows that linear Hawkes model with exponential decay function is well specified for SPY jumps. Another robustness check is a simulation study; we estimate linear Hawkes model with exponential decay function on data simulated from mixture of Poisson model. The simulation result suggests that linear Hawkes model may not be easily tricked; however, we cannot rule out the possibility that linear Hawkes model, when fitted on data generated by mixture of Poisson model, could produce a significant $CP > 0$.

Our conclusion and contributions are as follows: (1) both mixture of Poisson and linear Hawkes models suggest a simple Poisson process for price jumps should be rejected; (2) mixture of Poisson model has more precise dispersion ratios of price jumps than linear Hawkes model; (3) linear Hawkes model estimated on SPY jumps has stable and significant CP values given our decay functions. Overall, our results are consistent with Filimonov and Sornette (2015) that positive CP values given by linear Hawkes model must be interpreted with caution.

The paper is organized as follows. Section **2** gives a review on jump detection in high-frequency data and presents the jump test results. The phenomenon of over-dispersion is evident in daily counts of price jumps, as well as in their inter-arrival times. Section **3** describes the various specifications of mixture of Poisson and linear

Hawkes models. Section 4 reports the estimation results and compares the two models' performance in reproducing over-dispersion ratio. Section 5 provides a residual analysis on estimated Hawkes model and conducts a simulation study. Section 6 concludes. Technical details are provided in Appendix.

2. High-frequency Data and Jump Test

Jumps in high-frequency price data are identified by the test in Andersen, Bollerslev and Dobrev (2007), hereafter ABD test:

$$\frac{|r_{t,i}|}{\sqrt{(\eta_i BP_t)}} > \Phi_{1-\beta/2}, \tag{1}$$

where $r_{t,i}$ is the i-th return on day t, with $i = 1 \dots M$, $t = 1 \dots T$, $\sqrt{\eta_i BP_t}$ is the corresponding spot volatility estimates for $r_{t,i}$ and $\Phi_{1-\beta/2}$ is a critical value from standard normal N(0,1) distribution at significance level β[5]. The spot volatility estimate $\sqrt{\eta_{iBP_t}}$ is obtained with BP_t the Bi-Power Variation of day t, given by Barndorff-Nielsen and Shephard (2004, 2006):

$$BP_t = \frac{\pi}{2}\left(\frac{M}{M-1}\right)\sum_{i=1}^{M-1} |r_{t,i}|\,|r_{t,i+1}| \xrightarrow{as\, M\to\infty} \int_{t-1}^{t} \sigma^2(s)ds\,. \tag{2}$$

where M is the number of high-frequency interval per day and:

$$\eta_i = \frac{\sum_{t=1}^{T} r_{t,i}^2}{\sum_{t=1}^{T}\sum_{i=1}^{M} r_{t,i}^2}, \qquad i = 1 \dots M. \tag{3}$$

The coefficients η_i estimate the well-documented, U-shape intraday volatility pattern (IVP) across the sample and by definition, η_i sum to one, i.e. $\sum_{i=1}^{M} \eta_i = 1$.

The rationale of ABD jump test is straightforward: under the null hypothesis of no jumps in the underlying price process, a high-frequency return $r_{t,i}$ will be declared as containing jumps if after standardization it is too large to be afforded by a standard normal distribution. The Bi-Power Variation BP_t is a jump-robust estimate of the daily Integrated Variance $\int_{t-1}^{t} \sigma^2(s)ds$ for sufficiently large M, because as $M \to \infty$ the

[5] As the test will be applied to M intraday returns in a trading day, for a daily significance level α, the intraday significance level $\beta = 1 - (1-\alpha)^{1/M}$ to deal with the multiple comparison problem.

probability of seeing one price jump influencing both $r_{t,i}$ and $r_{t,i+1}$ is asymptotically zero. For a detailed review of ABD test see Dumitru and Urga (2012) and Tsai and Shackleton (2016).

The data used in this study is the traded prices of the ETF SPDR (ticker SPY), which is designed to track the performance of the S&P 500 index. The sample period is from January 02, 2002 to April 30, 2010; after a cleaning procedure is applied to the raw data, there are in total $T = 2076$ complete trading days[6]. Each day of 6.5 trading hours is then sampled every five and two minutes to give $M = 78$ and $M = 195$ respectively. The ABD test is then conducted for all the high-frequency, intraday returns in the SPY data. In the literature, two particular issues have been raised on the finite-sample performance of ABD test. First, when M is finite, there is a positive probability that consecutive price jumps may occur in two adjacent return intervals, hence jeopardizing the validity of BP_t; see Corsi, Pirino and Reno (2010). The second issue concerns the IVP estimates η_i; when return $r_{t,i}$ contains jumps in the first place, the estimates of η_i will be biased upward (Boudt, Cand Laurent 2011). In either case, the estimated values of spot volatility $\sqrt{(\eta_i BP_t}$ for $r_{t,i}$ will be over-estimated, which in turn will bias the ABD test toward a lower detection of price jumps.

In light of the finite-sample bias, the ABD test is modified as a three-step procedure. In the first step, the ABD test in (1) is applied to the SPY data; the first row of Table 1 shows there are 383 and 830 jumps detected at the 5-minute and 2-minute frequencies, under the significance level $\beta = 1 - (1 - \alpha)^{1/M}$, with daily significance level $\alpha = 1\%$[7]. In the next step, if on day t two consecutive returns $r_{t,i}$ and $r_{t,i+1}$ are detected as containing jumps, then BP_t will be re-calculated by replacing the quantity $|r_{t,i}||r_{t,i+1}|$ with $|r_{t,i-1}||r_{t,i+1}|$ and the jump test performed again. This step is repeated until all consecutive jumps on day t are adjusted with a replacement. The second row of Table 1 indicates that a few more jumps are identified as a result. Finally, jump-robust IVP coefficients η_i are obtained by replacing those $r^2_{t,i}$ which are classified as

[6]See Tsai and Shackleton (2016) for a description of SPY data.

[7]In ABD (2007), the jump test is implemented by assuming a constant intraday volatility level and thus they use a very conservative level $\alpha = 10^{-5}$. Since we specifically account for the U-shape pattern by η_i in (3) in our three-step procedure, we can choose a higher level $\alpha = 0.01$, which is shown to be reasonable in the extensive simulation study in Gilder, Shackleton and Taylor (2014) and Tsai and Shackleton (2016).

containing jumps with the average of $r_{t,i}^2$ over day t. In Table 1 it can be seen that more identifications of price jumps are made after this adjustment[8].

In total, there are 416 and 856 price jumps detected by the modified testing procedure. Table 2 reports the daily intensity and the summary statistics of the inter-arrival times between price jumps, which are measured in day[9]. On average 0.2004 and 0.4123 price jumps occur per day in the 5-minute and 2-minute SPY data. These daily counts of price jumps have variance 0.2278 and 0.4766 respectively. Therefore, the daily counts of price jumps display over-dispersion, with dispersion ratio 1.1367 and 1.1560, and thus it is not appropriate to use a simple Poisson model for the daily counts of price jumps. Over-dispersion can also be defined in terms of the inter-arrival times between price jumps, which have an exponential distribution if the arrivals of price jumps follow a Poisson process. In Table 2, the mean and standard deviation of the inter-arrival times show over-dispersion, with the dispersion ratios of both 5-minute and 2-minute frequencies being larger than one. Notably, the longest intervals between price jumps are more than 48 and 30 days at the 5-minute and 2-minute frequencies, respectively. Figure 1 plots the daily counts of 2-minute price jumps in the SPY data; there can be at most five jumps in a trading day and periods of zero count are visible.

3. Modelling Over-Dispersion in Price Jumps

To quantify the degree of over-dispersion in the jumps of SPY prices, two families of model are considered. The first is mixture of Poisson model, with the number of independent Poisson components $m = 2, 3$. The second is linear Hawkes model. Both models are capable of reproducing the over-dispersion phenomenon, but with different generating mechanisms. Mixture of Poisson model has intensity which alternates between m states, and thus jumps are more likely to occur during some periods than others. The linear Hawkes model stipulates that a subset of jumps forms a background Poisson process, which in turn will generate subsequent jumps in a self-exciting manner.

[8] We also consider replacing those $r_{t,i}^2$ which are classified as containing jumps with the median of $r_{t,i}^2$ over day t; the summary statistics of price jumps are similar and are reported in Table 2.

[9] Note that over-night period is ignored in the calculation of inter-arrival times.

3.1 Mixture of Poisson Model

Model Specification and Likelihood function

Let C_t denote the daily counts of price jumps which is assumed to have a mixture of Poisson distribution; the probability function of C_t is defined as follows (Zucchini and MacDonald 2010):

$$\Pr(C_t = c) = \sum_{j=1}^{m} \Pr(C_t = c|Y = j)\delta_j, \quad c = 0, 1, 2, \dots, \tag{4}$$

where Y is a mixing variable with m states and δ_j is the corresponding state probability summing up to one. In the case of $m = 2$, equation (4) becomes:

$$\Pr(C_t = c) = \delta_1 \frac{\lambda_1^c e^{-\lambda_1}}{c!} + (1 - \delta_1)\frac{\lambda_2^c e^{-\lambda_2}}{c!}, \tag{5}$$

which is a weighted average of two Poisson distributions with parameters λ_1 and λ_2. Since the individual Poisson components are independent, the mean and variance of C_t when $m = 2$ can be derived as:

$$\begin{aligned} \mathrm{E}[C_t] &= \delta_1\lambda_1 + (1 - \delta_1)\lambda_2, \\ \mathrm{Var}(C_t) &= \mathrm{E}[C_t] + \delta_1(1 - \delta_1)(\lambda_1 - \lambda_2)^2. \end{aligned} \tag{6}$$

Thus, the over-dispersion phenomenon can be generated because the daily counts of price jumps C_t has $\mathrm{Var}(C_t) > \mathrm{E}[C_t]$.

The likelihood function of a mixture of Poisson model with two independent components, given $c_1, \dots, c_T$ the T daily observed values of counts of price jumps, is given as:

$$L(\lambda_1, \lambda_2, \delta_1|c_1, \dots, c_T) = \prod_{t=1}^{T} \left[\delta_1 \frac{\lambda_1^c e^{-\lambda_1}}{c_t!} + (1 - \delta_1)\frac{\lambda_2^c e^{-\lambda_2}}{c_t!}\right], \tag{7}$$

For models with $m = 3$, the likelihood function can be expanded accordingly. Estimation of mixture models can be done with an EM algorithm[10]; for mixture of Poisson models considered in this study, direct maximization of log-likelihood function can

[10] See McLachlan and Peel (2000) and the review in Oliveira-Brochado and Martins (2005).

obtain satisfactory parameter values and their standard errors. Moreover, since the individual components are independent in mixture of Poisson models, it is expected that the unconditional intensity $\lambda = \sum_{t=1}^{T} c_t/T$ should be between the estimated values of λ and λ_2 from a mixture of Poisson model with $m = 2$, i.e. $\lambda_1^{m=2} < \lambda < \lambda_2^{m=2}$, provided that the mixture of two Poisson model has a significant improvement over the simple Poisson model.

Excessive Zeros and Zero-Inflated Poisson Model

In Table 2 the proportion of days with no jump identification is also reported; for the 5-minute SPY data 83.14% of 2076 days have no jumps, and for the 2-minute case it is 68.11%. In the literature, a zero-inflated Poisson (ZIP) model has been proposed to deal with the excessive observations of zeros in a count data; see the review by Ridout, Demétrio and Hinde (1998) and Rodriguez (2013):

$$\begin{cases} \Pr(C_t = 0) = w_z + (1 - w_z)e^{-\lambda} \\ \Pr(C_t = c) = (1 - w_z)\frac{\lambda^c e^{-\lambda}}{c!} \end{cases} \tag{8}$$

where w_z is the probability of “always zero” observation in the data. In a ZIP model for count data, the observed value 0 can come from two sources – one with probability w_z and the other with probability $(1 - w_z)$ from a Poisson model with parameter λ. Consequently, a ZIP model can be viewed as a special case of mixture of Poisson model by taking $\delta_1 = 1 - w_z$ and $\lambda_2 = 0$ in equation (5).

A further generalization can be made to construct a mixture of two Poisson models with zero-inflated characteristic, so the observed value 0 occurs with probability w'_z and $(1 - w'_z)$ from a mixture of two Poisson models:

$$\begin{cases} \Pr(C_t = 0) = w'_z + (1 - w'_z)\left[\delta_1 e^{-\lambda_1} + (1 - \delta_1)e^{-\lambda_2}\right] \\ \Pr(C_t = c) = (1 - w'_z)\left[\delta_1 \frac{\lambda_1^c e^{-\lambda_1}}{c!} + (1 - \delta_1)\frac{\lambda_2^c e^{-\lambda_2}}{c!}\right], \quad c = 1, 2, 3 \ldots \end{cases} \tag{9}$$

This model is termed zero-inflated mixture of Poisson (ZIMP) model, and will be considered in estimation.

3.2 Linear Hawkes Model

Given a point process of jump times $\{s_k\}, k = 1, \ldots, N(T)$, the linear Hawkes model of Hawkes (1971) specifies its conditional intensity $\lambda(s|\mathcal{F}_s)$ with the following self-exciting feature:

$$\lambda(s|\mathcal{F}_s) = b_0 + b_1 \int_0^s \omega(s-u)dN(u) = b_0 + b_1 \sum_{0 \le s_k < s} \omega(s - s_k), \tag{10}$$

where $b_0 > 0$ is a constant background intensity, $b_1 > 0$ is the increment of $\lambda(s|\mathcal{F}_s)$ at jump times $\{s_k\}$ and $\omega(\tau)$ is a non-increasing decay function of elapsed time $\tau_k = s - s_k \ge 0$. The conditional intensity begins at b_0 and increases by b_1 at the jump times $\{s_k\}$; between two jumps $\lambda(s|\mathcal{F}_s)$ declines at a rate controlled by the decay function $\omega(\tau)$. The linear Hawkes model can produce an expected value for the daily counts of price jumps C_t by integrating the conditional intensity:

$$\mathrm{E}[C_t] = \int_{t-1}^{t} \lambda(s|\mathcal{F}_s)ds.$$

When there are clustering in arrivals of price jumps, the conditional intensity $\lambda(s|\mathcal{F}_s)$ will reach to a high level; during a period of no jumps, however, it has a tendency to reduce to b_0. The linear Hawkes model therefore can reproduce over-dispersion phenomenon in the SPY price jumps data.

The vector of parameters $\boldsymbol{\theta}$ of linear Hawkes model can be estimated by the log-likelihood function given the observed jump times $\{s_k\}, k = 1, \ldots, N(T)$:

$$\ln L(\boldsymbol{\theta}) = \sum_{k=1}^{N(T)} \ln \lambda(s_k; \boldsymbol{\theta}) - \int_0^T \lambda(s; \boldsymbol{\theta})ds \tag{11}$$

The log-likelihood function is derived from the observed jump times $\{s_k\}, k = 1, \ldots, N(T)$, since "... <u>(these) points have occurred at a number of known locations, but nowhere else in the region. (p.127 Coles 2001)</u>." The maximum-likelihood estimates $\hat{\boldsymbol{\theta}}_{MLE}$ are shown to be asymptotically Gaussian under some regularity conditions (Ogata 1978). In particular, the following stationary condition must be maintained (Cox and Isham

1980; Vere-Jones & Ozaki 1982):

$$\mathrm{E}[\lambda(s)] = \mu \Rightarrow \frac{b_0}{1 - b_1 \int_0^\infty \omega(\tau)d\tau} = \mu \Rightarrow CP = b_1 \int_0^\infty \omega(\tau)d\tau < 1 \tag{12}$$

In Tsai (2017), the quantity $b_1 \int_0^\infty \omega(\tau)d\tau$ is termed clustering parameter (CP)[11]. A Poisson process by definition has $CP = 0$; consequently, a mixture of Poisson model with independent components also has $CP = 0$. On the other hand, if $CP = 1$, then the unconditional expectation of intensity $\mu \to \infty$, in which case the linear Hawkes process is considered un-stationary.

In the finance and seismology literature, there have been two popular choices for the decay function $\omega(\tau)$ in applying the Hawkes model to real data[12]:

$$\omega_{Exp}(\tau) = \exp(-\tau/a), \quad a > 0 \tag{13}$$

$$\omega_{I.P.}(\tau) = d^{(1+p)}(\tau + d)^{-(1+p)}, \quad c > 0. \tag{14}$$

Both functions begin with value one when $\tau = 0$ and approaches zero as $\tau \to \infty$ The $\omega_{Exp}(\tau)$ function represents an exponential rate of decay and hence a Markovian property in $\lambda(s|\mathcal{F}_s)$. The $\omega_{.P.}(\tau)$ function stands for an inverse-power rate of decay. Filimonov and Sornette (2015) point out that CP tends to be biased upward when the linear Hawkes model is equipped with the inverse-power decay function $\omega_{I.P.}(\tau)$. For example, an outlier event in $\{s_k\}$ with a large inter-arrival time $(s_k - s_{k-1})$ would be falsely recognized by $\omega_{I.P.}(\tau)$ as a long waiting time. The exponential decay function $\omega_{Exp}(\tau)$ on the contrary is less sensitive to the effect of outliers. Hardiman and Bouchaud (2014) also show that the estimated CP values are sensitive to the choice of decay function $\omega(\tau)$. In Table 3 we list the stationary condition $CP < 1$ for $\omega_{Exp}(\tau)$ and $\omega_{I.P.}(\tau)$, which will be maintained in the maximum log-likelihood estimation. We use $\exp(-\tau/a)$ instead of $\exp(-a/\tau)$ in function $\omega_{Exp}(\tau)$, and choose $p = 1$ in function $\omega_{I.P.}(\tau)$ to obtain CP as a product of two parameters, which entails the calculation of its standard error using the delta method[13]. The choice of $p \geq 1$ is also suggested by

[11]In Filimonov and Sornette (2015) and Ma and Chao (2016), it is called market endogeneity or reflexivity; in the literature of branching process, it is called the branching ratio.

[12]See the review in Ogata (1988), Bauwens and Hautsch (2009) and Law and Viens (2016).

[13]See Appendix for the derivation of the standard error of CP.

Filimonov and Sornette (2015) to avoid the case that "... events far in the past continue to influence the triggering of events far in the future." The two functions thus have the same number of parameters, but represent distinct rates of decay in the conditional intensity.

4. Estimation Results

In presenting the estimation results from the models in Section 3, it is useful to illustrate the relations among them:

$$\text{Poisson}\,(CP=0)\begin{cases}\text{Mixture of Poissons}\,(CP=0,\delta_j\neq 0))\\ \text{Linear Hawkes Model}\,(0<CP<1))\end{cases}$$

Note that the mixture of Poisson models will reduce to ZIP model if $m=2$, $\delta_1=1-w_z$ and $\lambda_2=0$.

4.1 ZIP and Mixture of Poisson Model

In Table 4, the estimation results of ZIP and mixture of Poisson models on 5-minute and 2-minute SPY price jumps are presented. The benchmark Poisson model has λ equal to the sample mean of daily counts in Table 2. This baseline model can be improved by introducing the probability of zero w_z as in the ZIP model. The estimated values of w_z are 0.4372 with standard error 0.0599 for the 5-minute SPY price jumps, and 0.2337 with standard error 0.0445 for the 2-minute jumps. The estimated values w_z are therefore highly significant, which renders the estimated values of λ increase to 0.3561 and 0.5381 respectively. The LR test also indicates that ZIP model outperforms the baseline Poisson model, with the LR test statistics 21.26 and 16.86 for the 5-minute and 2-minute SPY price jumps, hence rejecting the null of Poisson model with p-value $< 0.1\%$.

Mixture of Poisson models, on the other hand, show different performance for the SPY price jumps at the two frequencies. For the 5-minute SPY price jumps, mixture of Poisson models has the same maximized log-likelihood as ZIP model for all m. The estimated value of λ_1 is always close to 0.3561 which is obtained in the ZIP model. Thus, for the 5-minute jumps, mixture of Poisson model – with more parameters – does not outperform ZIP model. For the 2-minute SPY price jumps, however, mixture of Poisson models with $m = 2$ can give estimates of $\lambda_1 = 0.9144$ and $\lambda_2 = 0.2843$ with standard errors 0.2519 and 0.0585, respectively. Although the mixing parameter is estimated at $\delta_1 = 0.2032$ with standard error 0.1482, the likelihood ratio (LR) test statistic with respect to the benchmark Poisson model is 20.46, which is highly significant (p-value $<$ 0.1%) for a Chi-Square test with degree of freedom equal to 2. For mixture of Poisson models with $m = 3$, the maximized log-likelihood values remain at 1757.34.

Between ZIP and mixture of Poisson model, for 2-minute SPY price jumps, the model selection problem can be determined by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The values of AIC for the ZIP and mixture of Poisson model are 3522.28 and 3520.68; the values of BIC are 3531.78 and 3534.94, respectively. Thus, mixture of Poisson model with $m = 2$ is selected by AIC, whereas ZIP model is selected by BIC[14]. The choice of AIC versus BIC is beyond the scope of this paper[15], and in the subsequent analysis we will use both ZIP and mixture of Poisson model with $m = 2$ for 2-minute jumps. Our empirical evidence in Table 4 suggests that SPY price jumps do not follow a simple Poisson model, but can be described by mixture of Poisson models, which nest ZIP as a special case[16].

4.2 Results on Linear Hawkes Model

Table 5 reports the estimation results of linear Hawkes model for SPY jumps data. For both 5-minute and 2-minute jumps, the Hawkes model with $\omega_{Exp}(\tau)$ decay function can attain a higher log-likelihood value than with $\omega_{I.P.}(\tau)$ function. For 5-minute SPY jumps, the conditional intensity has a background intensity $b_0 = 0.1857$, compared with the daily intensity 0.2004 in Table 2. The clustering parameter CP is estimated at 0.0732. For 2-minute SPY jumps, the conditional intensity has a background intensity

[14] AIC$= 2(k - L_{max})$ and BIC$= (\ln n)k - 2L_{max}$, where k is the number of parameters in the model, n the sample size and L_{max} is the maximized log-likelihood.

[15] See, for example, Young (2005).

[16] Zero-inflated mixture of Poisson (ZIMP) models with $m \geq 2$ are also estimated but they have the same maximized log-likelihood as the ZIP model for the 5-minute SPY jumps, and also the same maximized log-likelihood as mixture of Poisson model with $m = 2$ for the 2-minute SPY jumps. To save space these results are not reported.

$b_0 = 0.3707$, compared with the daily intensity 0.4123 in Table 2. The clustering parameter CP is estimated at 0.1010.

One drawback of Hawkes model is that estimated CP values are sensible to the choice of decay functions (Hardiman and Bouchaud 2014 and Filimonov and Sornette 2015). However, in Table 5 the clustering parameter CP has similar values (0.0732 - 0.0892 for 5-minute and 0.1010 - 0.1184 for 2-minute jumps) between the two decay functions $\omega_{Exp}(\tau)$ and $\omega_{I.P.}(\tau)$, which have very different rates of decay. We also calculate the standard errors of CP using the asymptotic normality of the MLE parameters of Hawkes model (Ogata 1978). For 5-minute jumps, the standard errors of CP are 0.0172 and 0.0253 for the two decay functions; for 2-minute jumps they are 0.0140 and 0.0171 (see Appendix for calculation). These standard errors of CP indicate that the estimated CP values are robust. The results in Table 5 suggest that SPY price jumps do not follow a simple Poisson model, but display some degree of clustering as captured by the CP value of linear Hawkes model.

4.3 Comparison Using CP Approximation

In this section we compare the performance of mixture of Poisson and linear Hawkes models in describing the over-dispersion in the arrivals of SPY price jumps. We follow Hardiman and Bouchaud (2014) and consider their non-parametric estimator of clustering parameter $CP_{H.B.}$[17]:

$$CP_{H.B.} = 1 - \sqrt{\frac{\mathrm{E}[C_h]}{\mathrm{Var}[C_h]}}\ , \tag{15}$$

where $\mathrm{E}[C_h]$ and $\mathrm{Var}[C_h]$ are the expectation and variance of the number of counts in an interval h, which is one day in our setup. As a result, given an estimated CP value, we can obtain an approximated dispersion ratio by inverting (15). Table 6 reports the values of dispersion ratio for the models considered. For 5-minute jumps, the ZIP model has $\mathrm{E}[C_t] = 0.2004$ and $\mathrm{Var}[C_t] = 0.2316$ given by (6), and thus a dispersion ratio 1.1557. The linear Hawkes model with $\omega_{Exp}(\tau)$ function has an estimated $CP = 0.0732$, which gives a dispersion ratio 1.1642. These values are compared with the empirical

[17] See also p.170 of Bouchaud, Bonart, Donier and Gould (2018).

dispersion ratio 1.1367 reported in Table 2. For 2-minute jumps, the ZIP model has dispersion ratio 1.1258, whereas the mixture of Poisson model with $m = 2$ has dispersion ratio 1.1560, which matches the empirical value. On the other hand, the linear Hawkes model with $\omega_{Exp}(\tau)$ function has $CP = 0.1010$, which translates into a dispersion ratio of 1.2373.

The results in Table 6 indicate that linear Hawkes model has a stronger over-estimation on the dispersion ratio than mixture of Poisson (or ZIP) model. Note that from (12), we can obtain $\mathrm{E}[C_t] = b_0 + \mu CP$ for linear Hawkes model and these values can match the empirical ones. Thus, linear Hawkes model can reproduce the first moment of daily counts, but tend to overestimate the second moment. The mixture of Poisson model, with ZIP a special case, displays a better performance in reproducing the over-dispersion in SPY jumps.

5. Simulation and Residual Analysis

5.1 Simulation Study

To evaluate the performance of linear Hawkes model, mixture of Poisson model will be simulated with parameter values given in Table 4; linear Hawkes model will then be estimated on the simulated data. Mixture of Poisson model gives the probability of the number of price jumps per day, or C_t; linear Hawkes model on the other hand specifies the conditional intensity of jumps $\lambda(s|\mathcal{F}_s)$ which depends on the elapsed time since previous events. As a result, there will be two steps in the simulation:

Step 1: simulate two Poisson processes P_1 and P_2 over $(0, T]$ with daily intensities λ_1 and λ_2 such that the inter-arrival times $\{s_k - s_{k-1}\}, k = 1, 2, \ldots, N(T)$, follow exponential distribution with parameter $(1/\lambda_1)$ and $(1/\lambda_2)$ respectively;

Step 2: for each day t of unit length, determine if $u \sim U(0, 1) < \delta_1$; if the uniform random variable $u < \delta_1$, the first Poisson process P_1 is kept over day t, otherwise P_2 will prevail over day t. The resulted process is from a mixture of Poisson model.

In Step 1, it is not necessary to have the first jump time $s_0 = 0$, which can be done by setting a burn-in period for P_1 and P_2. In the simulation, this burn-in period has 100 days and will be discarded to obtain the simulated processes. A total of 100 processes from mixture of Poisson model are simulated using Step 1 and Step 2.

Table 7 reports the summary statistics of 100 simulation series and the results of linear Hawkes model estimated on them. In the left panel, when compared with 5-minute (416) and 2-minute SPY jumps (856), the simulated data has average daily counts 416.85 and 854.20[18]. The simulated data also has average variance of daily counts and average percentage of zeros in daily counts matching empirical values. For the inter-arrival times between jumps, the simulated data has on average 4.9792 and 2.4298 days for 5-minute and 2-minute jumps, compared with 4.9656 and 2.4206 from real data. However, the simulated data sees a lower standard deviation (5.3457 and 2.5749 days) of inter-arrival times, compared with 6.2241 and 3.2741 from real data. Consequently, the second dispersion ratio calculated from simulated data is lower than those of SPY jumps. This discrepancy between simulated and real data should be considered in interpreting the following estimation results on linear Hawkes model[19].

In the right panel of Table 7, the reported are parameter values of Hawkes model estimated on the simulated data, which are averages over 100 simulations. When compared with those from SPY jumps, the only parameter that remains close is b_0: 0.1842 v.s. 0.1857 for 5-minute and 0.3793 v.s. 0.3707 for 2-minute jumps. The values of b_1 and a are very different from those estimated from real data – the average of b_1 is smaller (0.1808 v.s. 1.4299 and 0.1760 v.s. 4.1982) albeit significant, whereas the average of a is larger (0.5513 v.s. 0.0512 and 0.5937 v.s. 0.0241) but insignificant. These changes of parameter values may indicate that the simulated data has a different structure or dynamics than SPY price jumps. Finally, although the estimated CP values 0.0841 and 0.0748 from simulated data are comparable to those from real data, their standard errors 0.0401 and 0.0367 are not very small and we cannot reject $CP = 0$ easily.

[18] The summary statistics are obtained as averages over 100 simulated series.

[19] This discrepancy may be reduced by designing a more sophisticated simulation algorithm for mixture of Poisson model; we will leave this task as future research.

From the simulation, we may conclude that linear Hawkes models, when fitted on data generated by mixture of Poisson models, can have a positive but insignificant CP value. However, this result could be affected by the lower standard deviation of inter-arrival times in our simulated data.

5.2 Residual Analysis of Linear Hawkes Model

Both functions begin with value one when, linear Hawkes model "…comes closest to fulfilling, for point processes, the kind of role that the autoregressive model plays for conventional time series." Indeed, the two decay functions $\omega_{Exp}(\tau)$ and $\omega_{(I.P.)(\tau)}$ mirror the choice between short memory (exponential decay) and long memory (inverse-power decay). For linear Hawkes model, there is also a goodness-of-fit test that resembles those in time-series models – known as Random Time Change – which is due to Papangelou (1972) and Ogata (1988). Specifically, the following transformed event times $\mathcal{S}_k$ from jump times s_k will form a Poisson process with unit intensity[20]:

$$\Lambda(s_k) = \int_0^{s_k} \lambda(s|\mathcal{F}_s)ds = S_k.$$

Ogata (1988) then develops a residual analysis by exploiting the fact that when $\lambda(s|\mathcal{F}_s)$ is not correctly specified, the realizations of $\{S_k\}_{k=1...N(T)}$ will deviate from the unit-intensity Poisson process. The evaluation of linear Hawkes model $\lambda(s|\mathcal{F}_s)$ thus amounts to testing for randomness in the residual process S_k.

In Figure 2 we plot the number of detected jumps k (vertical axis) against the residuals S_k (horizontal axis). If $\{S_k\}_{k=1...N(T)}$ follows a unit-rate Poisson process, the plot should approximate a 45-degree straight line from the origin to $N(T)$. For 5-minute jumps $N(T) = 416$, and for 2-minute jumps $N(T) = 856$. In both figures the empirical plots do not deviate from the 45-degree line. A KS (Kolmogorov-Smirnov) test confirms this result, with KS statistic equal to 0.0024 and 0.0012 respectively, which are smaller than the critical value 0.0662 and 0.0462 at 5% significance level. Hence, the linear Hawkes model with exponential decay function is considered well-specified for the SPY price jumps.

[20] The reverse result is also valid; for general s, $\Lambda(s)$ is defined to be the compensator of the counting process $N(s)$ such that $N(s) - \Lambda(s)$ is a martingale with respect to the history of the point process.

6. Conclusion

In this study we compare two methods in estimating the degree of over-dispersion in SPY price jumps. The first is mixture of Poisson model, which by construction has a clustering parameter $CP = 0$. The second is linear Hawkes model, which has $CP > 0$. We consider various specifications within each class of model; we then compare the dispersion ratios generated by these models and see if they match the empirical values. Our results show that both models should be preferred over the simple Poisson model, with mixture of Poisson model – which nests ZIP as special case – giving more precise dispersion ratios for SPY jumps than linear Hawkes model. We also make contributions to the literature firstly by obtaining stable CP values in linear Hawkes models with different decay functions, and secondly by inferring the dispersion ratio of data using the non-parametric method of Hardiman and Bouchaud (2014). Moreover, we provide the analytic and empirical values of standard errors of CP which are afforded by the two decay functions considered.

We conduct a simulation study and estimate linear Hawkes model on data generated from mixture of Poisson model. The results cannot rule out the possibility that in such a case, a linear Hawkes model would give a $CP > 0$. We also validate the specification of linear Hawkes model by implementing the residual analysis developed by Ogata (1988). In general, our findings are consistent with Filimonov and Sornette (2015) that results estimated from linear Hawkes models on high-frequency data should be analysed with caution. In particular, price jumps are obtained with detection errors whose distribution is unknown. To further investigate if true price jumps have $CP > 0$, more sophisticated methods are required such as a Hawkes model with state-dependent background intensity or a hidden Markov Poisson model (Poisson-HMM), as in Zucchini and Macdonald (2009). We will leave these tasks in future research.

Acknowledgments

I thank two anonymous referees and the editor for providing very useful comments which significantly improve the content of paper. I also benefit a lot from discussions with Didier Sornette and Vladimir Filimonov. All errors remain mine.

Appendix. Calculating the Standard Error of CP

The clustering parameter CP is given as a product of two parameters in the linear Hawkes models considered in Table 3. Ogata (1978) shows that the MLE estimates of linear Hawkes model are asymptotically normal, under some regularity conditions. We therefore consider the variance of the product of two normal variables X and Y which are correlated by a correlation coefficient ρ:

$$X = \mu_X + \sigma_X Z \sim N(\mu_X, \sigma_X^2)$$

$$Y = \mu_Y + \sigma_Y(\sigma Z + \sqrt{1-\rho^2}W) = \mu_Y + \sigma_Y U \sim N(\mu_Y, \sigma_Y^2)$$

where Z and W are independent standard normal random variables. Consequently: Also, $\mathrm{Cov}(Z, U^2) = 0$ since $\mathrm{Cov}(Z, Z^2) = 0$ and $\mathrm{Cov}(Z, ZW) = 0$. Likewise, $\mathrm{Cov}(Z^2, U) = 0$ and so:

$$\mathrm{Cov}(Z^2, U^2) = \rho^2 \mathrm{Var}(Z^2) = \rho^2(\mathrm{E}[Z^4] - 1) = 2\rho^2$$

Then we have

$$\begin{aligned}
\mathrm{Var}(XY) &= \mathrm{E}[X^2Y^2] - (\mathrm{E}[XY])^2 \\
&= \mathrm{E}[X^2]\mathrm{E}[Y^2] + \mathrm{Cov}(X^2, Y^2) - (\mathrm{E}[X]\mathrm{E}[Y] + \mathrm{Cov}(X, Y))^2 \\
&= (\mu_X^2 \sigma_X^2)(\mu_Y^2 + \sigma_Y^2) + \mathrm{Cov}(X^2, Y^2) - (\mu_X\mu_Y + \rho\sigma_X\sigma_Y)^2,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{Cov}(X^2, Y^2) &= \mathrm{Cov}(\mu_X^2 + 2\mu_X\sigma_X Z + \sigma_X^2 Z^2, \mu_Y^2 + 2\mu_Y\sigma_Y U + \sigma_Y^2 U^2) \\
&= \mathrm{Cov}(2\mu_X\sigma_{sX} Z + \sigma_X^2 Z^2, 2\mu_Y\sigma_Y U + \sigma_Y^2 U^2) \\
&= 4\mu_X\mu_Y \mathrm{Cov}(X, Y) + 2[\mathrm{Cov}(X, Y)]^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{Var}(XY) &= (\mu_X^2 + \sigma_X^2)(\mu_Y^2 + \sigma_Y^2) + 4\mu_X\mu_Y \mathrm{Cov}(X, Y) + 2[\mathrm{Cov}(X, Y)]^2 \\
&\quad - \mu_X^2\mu_Y^2 - 2\mu_X\mu_Y \mathrm{Cov}(X, Y) - [\mathrm{Cov}(X, Y)]^2 \\
&= \sigma_X^2\sigma_Y^2 + \mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2 + 2\mu_X\mu_Y \mathrm{Cov}(X, Y) + [\mathrm{Cov}(X, Y)]^2 \qquad \text{(A.1)}
\end{aligned}$$

In calculating the variance of $CP = b_1 a$ or $CP = b_1 d$ by equation (A.1), the values of μ_X and μ_Y are given by the MLE estimates of (b_1, a) or (b_1, d) in Table 5. The values of σ_X^2, σ^2 and $\mathrm{Cov}(X, Y)$ are given by the variance-covariance matrix which is obtained from the inverse of Hessian function as in Table A1 below.

Table A1: Empirical Values of Inverse of Hessian Function.

		2-min SPY jumps		
$\omega_{Exp}(\tau)$	b_0	1.86E-4	7.83E-4	-1.02E-5
	b_1	–	0.6634	-0.0031
	a	–	–	2.35E-5
$\omega_{I.P.}(\tau)$	b_0	1.57E-4	6.43E-4	-9.18E-6
	b_1	–	1.4655	-0.0048
	d	–	–	2.29E-5
		5-min SPY jumps		
$\omega_{Exp}(\tau)$	b_0	8.55E-5	9.98E-5	-1.88E-5
	b_1	–	0.1733	-0.0036
	a	–	–	1.60E-4
$\omega_{I.P.}(\tau)$	b_0	9.43E-5	3.84E-4	-4.15E-5
	b_1	–	0.2783	-0.0092
	d	–	–	4.79E-4

Table 1: ABD Jump Test Results, 2002 – 2010, SPY.

ABD Test: $|r_{t,i}|/\sqrt{BP_t \eta_i} > \Phi_{1-\beta/2}$

$\beta = 1 - (1-\alpha)^{1/M}$, $\alpha = 1\%$

Frequency	5m	2m
No. of Jumps, no adjustment	383	830
No. of Jumps - adjusted for consecutive jumps	384	839
No. of Jumps - adjusted for consecutive jumps and the effect of jumps in IVP	416	856

Table 1 reports the number of price jumps detected in high-frequency SPY data, when there is no adjustment and when two adjustments are made for consecutive jumps and the effect of jumps in intraday volatility pattern (IVP). The test is performed with daily significance level $\alpha = 1\%$ and $\beta = 1 - (1-\alpha)^{1/M}$.

Table 2: Empirical Values of Inverse of Hessian Function.

Frequency	5-min	2-min	5-min	2-min
Jump-robust η_i	Replaced by average of $r_{t,i}^2$ on day t		Replaced by median of $r_{t,i}^2$ on day t	
Daily Count of Jumps				
Total	416	856	417	860
Daily Intensity	0.2004	0.4123	0.2009	0.4143
Variance	0.2278	0.4766	0.2281	0.4789
Dispersion	1.1367	1.1560	1.1354	1.1561
% of Zero	83.14%	68.11%	83.09%	68.02%
Inter-Arrival Times				
Mean	4.9656	2.4206	4.9677	2.4093
Std. Dev.	6.2241	3.2741	6.2168	3.2597
Maximum	48.1154	30.4462	48.1154	30.4462
Minimum	0.0128	0.0051	0.0128	0.0051
Dispersion	1.2534	1.3526	1.2514	1.3530

Table 2 reports summary statistics of price jump data when two adjustments are made. Daily Intensity is defined as the ratio of number of detected jumps over the number of days in the SPY sample. Over-Dispersion can be seen for the daily numbers of jumps, as well as in the inter-arrival times between jumps. We consider two different methods in the jump-robust IVP factor η_i; the summary statistics are fairly similar for both daily count of jumps and their inter-arrival times.

Table 3: Stationary Conditions for Linear Hawkes Process.

$\omega(\tau)$, $\tau \geq 0$	$\int \omega(\tau)d\tau$	$CP = b_1 \int_0^\infty \omega(\tau)d\tau < 1$
$\omega_{Exp}(\tau) = \exp(-\tau/a)$	$-ae^{-\tau/a}$	$b_1 a < 1$
$\omega_{I.P.}(\tau) = d^2(\tau + d)$	$-d^2(\tau + d)^{-1}$	$b_1 d < 1$

Table 3 lists the stationary conditions for the two decay functions in linear Hawkes model. The stationary conditions are maintained in the maximum likelihood estimation. We use $\exp(-\omega/a)$ instead of $\exp(-ar)$ in function $\omega_{Exp}(\tau)$, and choose $p = 1$ in function $\omega_{I.P.}(\tau)$ to obtain CP as a product of two parameters, which entails the calculation of its standard error using the delta method.

Table 4: Estimation of ZIP and Mixture of Poisson Model.

	5-min SPY jumps				*2-min SPY jumps*			
Model	$m=1$	ZIP	$m=2$	$m=3$	$m=1$	ZIP	$m=2$	$m=3$
δ_1	–	–	0.5628	0.2529	–	–	0.2032	0.5435
	–	–	(0.1397)	(0.1561)	–	–	(0.1482)	(0.0031)
δ_2	–	–	–	0.4372	–	–	–	0.2010
	–	–	–	(0.3268)	–	–	–	(0.2770)
λ_1	0.2004	0.3561	0.3560	0.3561	0.4123	0.5381	0.9144	0.9169
	(0.0098)	(0.0414)	(0.0592)	(0.1595)	(0.0141)	(0.0365)	(0.2519)	(0.4965)
w_z	–	0.4372	–	–	–	0.2337	–	–
	–	(0.0599)	–	–	–	(0.0445)	–	–
λ_2	–	–	0.0000	0.0000	–	–	0.2843	0.2730
	–	–	(0.0492)	(0.1273)	–	–	(0.0585)	(0.1639)
λ_3	–	–	–	0.3560	–	–	–	0.3117
	–	–	–	(0.1770)	–	–	–	(0.2017)
$-\log L$	1132.09	1121.46	1121.46	1121.46	1767.57	1759.14	1757.34	1757.34
LR test X^2	–	21.26*	21.26*	–	–	16.86*	20.46*	–
AIC	2266.18	2246.92*	2248.92	2250.92	3537.14	3522.28	3520.68*	3522.68
BIC	2270.21	2254.98*	2261.01	2267.04	3541.89	3531.78*	3534.94	3541.69

Table 4 reports the estimation of ZIP and mixture of Poisson models for the 5-minute and 2-minute SPY price jumps. The baseline Poisson model has $\lambda = 0.2004$ and 0.4123 which are the sample mean of daily counts in Table 2. For 5-min jumps, both AIC and BIC select the ZIP model. For 2-min jumps, mixture of Poisson model with $m = 2$ is selected by AIC, whereas ZIP model is selected by BIC. AIC$= 2(k - L_{max})$ and BIC$= (\ln n)k - 2L_{max}$, where k is the number of parameters in the model, n the sample size and L_{max} is the maximized log-likelihood. The likelihood ratio (LR) test is against the benchmark model ($m = 1$) and is all significant with small p values.

Table 5: Estimation of Linear Hawkes Model on SPY Price Jumps.

	5-min SPY jumps		*2-min SPY jumps*	
Model	$\omega_{Exp}(\tau)$	$\omega_{I.P.}(\tau)$	$\omega_{Exp}(\tau)$	$\omega_{I.P.}(\tau)$
b_0	0.1857	0.1825	0.3707	0.3635
	(0.0092)	(0.0097)	(0.0136)	(0.0125)
b_1	1.4299	1.5029	4.1982	5.6127
	(0.4163)	(0.5275)	(0.8145)	(1.2106)
a	0.0512	–	0.0241	–
	(0.0127)	–	(0.0048)	–
d	–	0.0593	–	0.0211
	–	(0.0219)	–	(0.0048)
$-\log L$	1057.52*	1059.91	1513.97*	1515.29
CP	0.0732	0.0892	0.1010	0.1184
$S.E.(CP)$	0.0172	0.0253	0.0140	0.0171

Table 5 reports the estimation of linear Hawkes model for SPY jumps data. For both 5-minute and 2-minute jumps, the Hawkes model with $\omega_{Exp}(\tau)$ decay function attains a higher log-likelihood than $\omega_{I.P.}(\tau)$ function. The clustering parameter CP has similar values (0.0732 - 0.0892 for 5-minute and 0.1010 - 0.1184 for 2-minute jumps) between the two decay functions $\omega_{Exp}(\tau)$ and $\omega_{I.P.}(\tau)$. We also calculate the standard errors of CP (see Appendix for calculation); these standard errors indicate that the estimated CP values are robust.

Table 6: Comparison of Mixture of Poisson and Linear Hawkes Model.

	5-min SPY jumps			*2-min SPY jumps*			
	Data Values	MixPoi ZIP	Hawkes $\omega_{Exp}(\tau)$	Data Values	MixPoi ZIP	MixPoi $m=2$	Hawkes $\omega_{Exp}(\tau)$
Daily Intensity $\mathrm{E}[C_t]$	0.2004	0.2004	0.2004	0.4123	0.4123	0.4123	0.4123
Variance $\mathrm{Var}[C_t]$	0.2278	0.2316	–	0.4766	0.4642	0.4766	–
Estimated CP	–	–	0.0732	–	–	–	0.1010
Dispersion Ratio	1.1367	1.1557	1.1642	1.1560	1.1258	1.1560	1.2373

Table 6 compares linear Hawkes model with mixture of Poisson model in reproducing the over-dispersion in SPY jumps. For 5-minute jumps, the ZIP model has $\mathrm{E}[C_t] = 0.2004$ and $\mathrm{Var}[C_t] = 0.2316$ given by (6), and thus a dispersion ratio 1.1557. The linear Hawkes model with $\omega_{Exp}(\tau)$ function has an estimated $CP = 0.0732$, which gives a dispersion ratio 1.1642 by inverting (17). For 2-minute jumps, the ZIP model has dispersion ratio 1.1258, whereas the mixture of Poisson model with $m = 2$ has dispersion ratio 1.1560, which matches the empirical value. The linear Hawkes model with $\omega_{Exp}(\tau)$ function has $CP = 0.1010$, which translates into a dispersion ratio of 1.2373.

Table 7: Estimation of Linear Hawkes Model on Data Simulated from Mixture of Poisson Model.

Summary Statistics			Hawkes Model with $\omega_{Exp}(\tau)$		
	5-min SPY Jumps	*Simulated Data*		*5-min SPY Jumps*	*Simulated Data*
Count	416	416.85	b_0	0.1857	0.1842
Daily Intensity	0.2004	0.2008		(0.0092)	(0.0117)
Variance	0.2278	0.2306	b_1	1.4299	0.1808
Dispersion 1	1.1367	1.1481		(0.4163)	(0.0774)
% of Zero	83.14%	83.06%	a	0.0512	0.5513
Mean	4.9656	4.9792		(0.0127)	(0.3132)
Std. Dev.	6.2241	5.3457	CP	0.0732	0.0841
Dispersion 2	1.2534	1.0736		(0.0172)	(0.0401)
	2-min SPY Jumps	*Simulated Data*		*2-min SPY Jumps*	*Simulated Data*
Count	856	854.20	b_0	0.3707	0.3793
Daily Intensity	0.4123	0.4115		(0.0136)	(0.0177)
Variance	0.4766	0.4731	b_1	4.1982	0.1760
Dispersion 1	1.1560	1.1495		(0.8145)	(0.0744)
% of Zero	68.11%	68.10%	a	0.0241	0.5937
Mean	2.4206	2.4298		(0.0048)	(0.3415)
Std. Dev.	3.2741	2.5749	CP	0.1010	0.0748
Dispersion 2	1.3526	1.0597		(0.0140)	(0.0367)

Table 7 reports the estimation results of linear Hawkes model on simulated data from mixture of Poisson model. In the left panel, the simulated data has roughly the same daily counts and percentage of zeros as real data, except for the smaller standard deviation of inter-arrival times. Consequently, the second dispersion ratio calculated from simulated data is lower than those of SPY jumps. In the right panel, the reported Hawkes parameter values from the simulated data are different from those given by real data, except for the background intensity b_0. The estimated CP values 0.0841 and 0.0748 from the simulated data are comparable to those from real data, but their standard errors 0.0401 and 0.0367 are not small enough to confidently reject $CP = 0$.

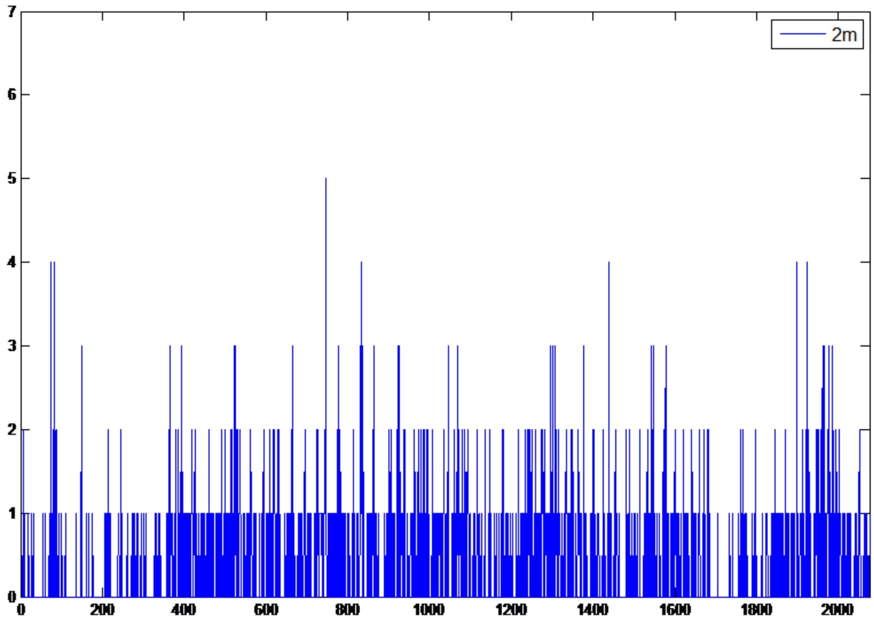
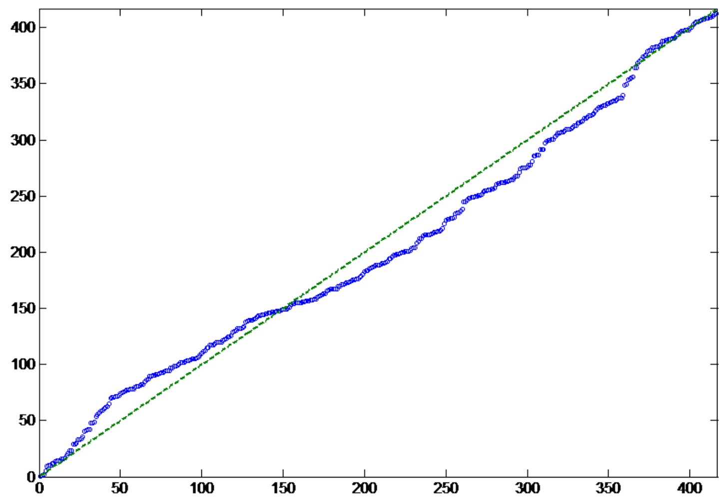


Figure 1 plots the time series of counts on jump detections for 2-minute SPY returns during 2002 – 2010. There are many daily observations equal to zero in the sample.

Figure 1: Daily Counts on Jump Detections, SPY.



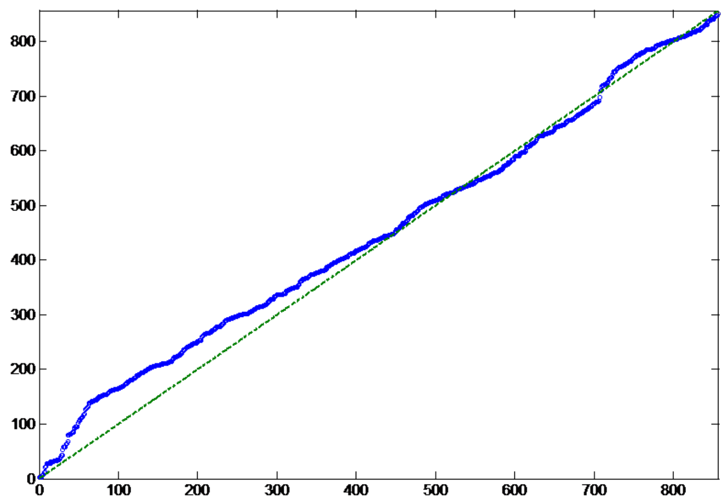


Figure 2 plots the number of jumps against the transformed time $S_k = \int_0^{s_k} \lambda(s|\mathcal{F}_s)ds$ for SPY jumps $\{s_k\}, k = 1 \dots N(T)$, with the conditional intensity $\lambda(s|\mathcal{F}_s)$ given by the linear Hawkes model with exponential decay function in Table 5. When $\lambda(s|\mathcal{F}_s)$ is correctly specified, the realizations of $\{s_k\}, k = 1 \dots N(T)$ will form a Poisson process with unit rate, and so the plot should approximate a 45-degree straight line from the origin to $N(T)$. For 5-minute jumps $N(T) = 416$, and for 2-minute jumps $N(T) = 856$. In both figures the empirical plots do not deviate from the 45-degree line. A KS (Kolmogorov-Smirnov) test confirms this result, with KS statistic equal to 0.0024 and 0.0012 respectively, which are smaller than the critical value 0.0662 and 0.0462.

Figure 2: Residual Analysis of Linear Hawkes model.

References

[1] Andersen T.G., T. Bollerslev and D. Dobrev (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, pages 125-180.

[2] Barndorff-Nielsen, O.E. & N. Shephard. 2004, Power and bi-power variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2, pages 1-48.

[3] Barndorff-Nielsen O.E. & N. Shephard. 2006, Econometrics of testing for jumps in financial economics using bi-power variation. *Journal of Financial Econometrics*, 4, pages 1-30.

[4] Bauwens L. and N. Hautsch (2009). Modelling financial high frequency data using point processes, in Handbook of Financial Time Series, Springer.

[5] Bouchaud, J.-P., J. Bonart, J. Donier and M. Gould (2018), Trades, Quotes and Prices: Financial Markets Under the Microscope, Cambridge University Press.

[6] Boudt, K., C. Croux and S. Laurent (2011). Robust estimation of intraweek periodicity in volatility and jump detection. *Journal of Empirical Finance*, 18, pages 353-367.

[7] Coles, S. (2001) An Introduction to Statistical Modelling of Extreme Values, Springer, London.

[8] Corsi, F., D. Pirino and R. Renò (2010). Threshold bi-power variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159, pages 276-288.

[9] Cox, D.R. and V. Isham (1980). Point Processes. Chapman & Hall.

[10] [10]Daley D.J. and D. Vere-Jones (2003). An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods, Springer.

[11] Dumitru, A.-M. and G. Urga (2012). Identifying jumps in financial assets: a comparison between nonparametric jump tests. *Journal of Business and Economics Statistics*, 30, pages 242-255.

[12] Eraker B., M. Johannes & N. Polson. (2003) The impact of jumps in volatility and returns. *Journal of Finance*, 58, pages 1269-1300.

[13] Ferriani, F. and P. Zoi (2020) The dynamics of price jumps in the stock market: an empirical study on Europe and U.S., The European Journal of Finance, DOI: 10.1080/1351847X.2020.1740288.

[14] Filimonov V. and D. Sornette (2015). Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15, pages 1293-1314.

[15] Filimonov V., S. Wheatley and D. Sornette (2015). Effective measure of endogeneity for the Autoregressive Conditional Duration point processes via mapping to the self-excited Hawkes process. *Communications in Nonlinear Science and Numerical Simulation*, 22, pages 23-37.

[16] Fulop, A., J. Li and J. Yu (2015). Self-exciting jumps, learning, and asset pricing implications. *Review of Financial Studies*, 28, pages 876-912.

[17] Giles, D. (2010). Notes on the zero-inflated Poisson Regression model, Department of Economics, University of Victoria.

[18] Hainaut D. and F. Moraux (2018) Hedging of options in the presence of jump clustering. *Journal of Computational Finance*, 22, pages 1-35.

[19] Hardiman, S.J., N. Bercot and J.-P. Bouchaud (2013). Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B*, 86, page 442.

[20] Hardiman, S.J. and J.-P. Bouchaud (2014). Branching ratio approximation for the self-exciting Hawkes process. *Physical Review E*, 90, 062807.

[21] Hawkes, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58, pages 83-90.

[22] Law, B. and F. G. Viens. 2016. Hawkes processes and their applications to high frequency data modelling, in Handbook of High-frequency Trading and Modelling in Finance (ed. I. Florescu, M. C. Mariani, H. E. Stanley and F. G. Viens), pages 183-212, Hoboken: Wiley.

[23] Ma T. and Y.H. Chao (2016). A new measure of market fragility – can trading network structure explain market reflexivity? Working paper, National Sun Yat-sen University.

[24] Maneesoonthorn, W., C. S. Forbes and G. M. Martin (2017) Inference on self-exciting jumps in prices and volatility using high-frequency measures. *Journal of Applied Econometrics*, 32, pages 504-532.

[25] McLachlan, G.J. and D. Peel (2000). Finite Mixture Models, Wiley, New York.

[26] Oliveira-Brochado A. and F. V. Martins (2005). Assessing the number of components in mixture models: A review, Universidade do Porto working paper.

[27] Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83, pages 9-27.

[28] Papangelou, F. (1972) Integrability of expected increments of point processes and related random change of scale. *Transactions of the American Mathematical Society*, 165, pages 483-506.

[29] Pesaran, M.H. and M. Weeks (2003). Non-nested hypothesis testing: An overview, in A Companion to Theoretical Econometrics, B.N. Baltagi, Ed., Blackwell, Oxford, pages 279-309.

[30] Ridout, M., C.G.B. Demetrio and J. Hinde (1998). Models for count data with many zeros, International Biometric Conference, Cape Town.

[31] Rodríguez, G. (2013). Lecture notes on Generalized Linear Models – models

for count data with over-dispersion. URL: http://data.princeton.edu/wws509/notes/

[32] Tsai, P.C. and M.B. Shackleton (2016). Detecting jumps in high-frequency prices under stochastic volatility: a data-driven approach, in Handbook of High-frequency Trading and Modelling in Finance. John Wiley.

[33] Tsai, P.C. (2017). Do jumps in financial prices cluster? Evidence from high-frequency data, 2017 Asian Meeting of Econometric Society, Hong Kong.

[34] Vere-Jones, D. and T. Ozaki (1982). Some examples of statistical inference applied to earthquake data. *Annals of the Institute of Statistical Mathematics*, 34, pages 189-207.

[35] Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica*, 57, pages 307-333.

[36] Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92, pages 937-950.

[37] Zucchini, W. and I.L. Macdonald (2009). Hidden Markov Models for Time Series: An Introduction Using R, Chapman & Hall/CRC.

[Received January 2020; accepted May 2021.]

關於價格跳躍的過度離散程度: 比較混合 Poisson 與線性 Hawkes 模型的實證研究

蔡秉真†

國立中山大學財務管理學系

摘　要

價格跳躍可能會有群聚發生的特性, 而這一性質通常以在計數資料中常見的過度離散的形式所表現出來, 即計數資料的時間序列之變異數將大於其平均數, 因此使用一簡單的 Poisson 模型不足以描述此性質。本文探討了兩方法 -混和 Poisson 模型與線性 Hawkes 模型 – 兩者均能產生過度離散性, 但前者的群聚係數爲 0, 而後者的群聚係數大於 0。兩模型的不同版本都配適到由高頻資料得出的價格跳躍數據中, 進而比較所得出的過度離散估計值。模擬實驗與殘差分析也提供了所得出結果的可靠性。

關鍵詞: 價格跳躍、過度離散、混合 Poisson 模型、線性 Hawkes 模型。

JEL classification: C13, G10.

†通訊作者: 蔡秉真
E-mail: vincenttsai@g-mail.nsysu.edu.tw