Probability of Informed No-Tradings: A Copula-Based PIN Model with Zero-Inflated Poisson Distributions

Chu-Lan Michael Kao[1,*], *Emily* Lin[2,**], and *Shan-Chi* Wu[3,***]

[1]Institue of Statistics, National Yang Ming Chiao-Tung University, Taiwan
[2]Department of Fashion Administration and Management, St. John's University, Taiwan
[3]Institute of Statistical Science, Academia Sinica, Taiwan

Abstract. Classical probability of informed trading (PIN) models assume that, given the information scenario, the number of buy and sell order flows are independently Poisson distributed, which imposes an assumption on the probability of no-trades. However, empirical data shows that the implied probabilities of no-trades do not match the aforementioned Poisson and independent assumptions. Therefore, we propose a new PIN model that better fits the data by using zero-inflated Poisson distributions and copula functions, which allow us to match the probability of no-trades. The expectation conditional maximization (ECM) is further proposed to tackle the parameter fittings, which is verified by simulation studies. The empirical studies show that this model outperforms the original PIN models, with significant parameters on the zero-inflations as well as copulas. In particular, we find that it is possible for an information to simultaneously increase the probability of no trade and boost up the average number of transactions, which contradicts the intuition.

1 Introduction

In finance literature, the probability of informed trading (PIN) models provide frameworks to study the flow of buys and sells in the market. In its simplest form, PIN models assume that, for each time period, given the type of information in the market, the number of buys and sells in that period follows a joint independent Poisson distribution, independent to the flows in other time periods. For example, [1] consider a market with good, bad, or no private information, and assume that the joint order flows follow the aforementioned distributions with different parameters. For more complicated PIN models, see [2] for instance.

However, the independence assumption used in these existing models is questionable, as it basically assumes that, given the market information, there is no interaction between the buys and sells, which is quite unrealistic. Indeed, [3] has verified this concern using empirical data, showing that the independence assumption is violated in the real world.

To show this violation more clearly, we give a simple example to illustrate the correlation between the buy and the sell in the data. Figure 1 shows the record of buy and sell transaction counts at 5-minute intervals for the stock AAPL on January 2nd, 2002. It is clear that the

*e-mail: chulankao@gmail.com
**e-mail: mlin@mail.sju.edu.tw
***e-mail: 594537@gmail.com

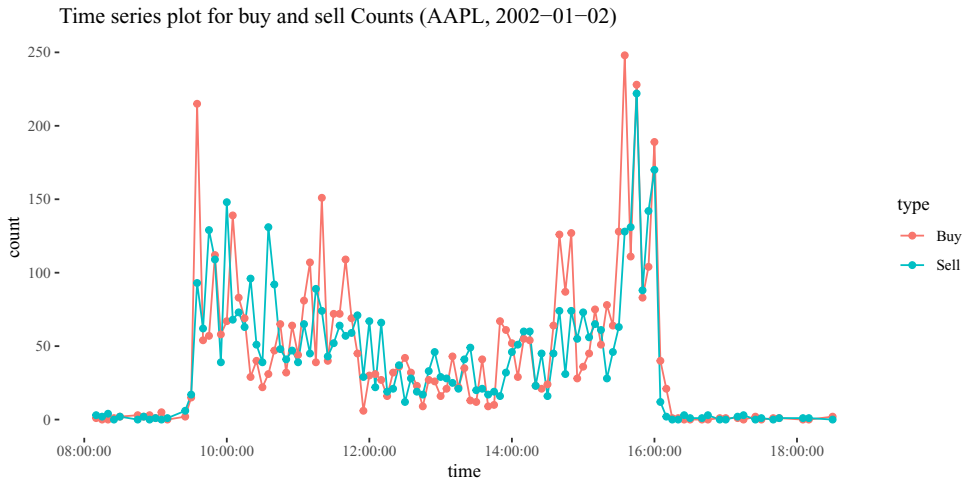


Figure 1: Times series of buys and sells (AAPL, 2002-01-02). Each point is the record of buy and sell transaction counts at 5-minute intervals. The buys and sells patterns are correlated.

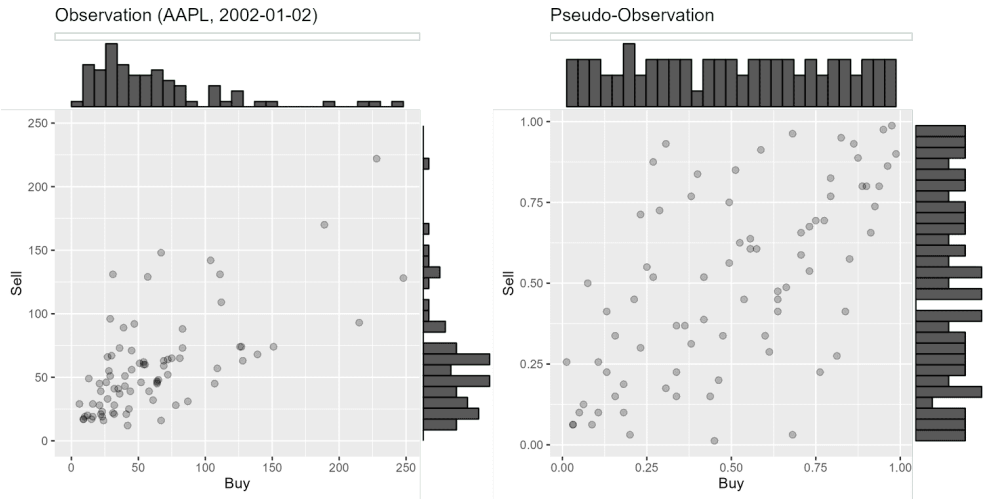


Figure 2: The scatter plot for buys and sells (AAPL, 2002-01-02). Left panel: original data. Right panel: pseudo-observation. The data points in the upper left and lower right corners are relatively sparse, so the data is positively correlated.

patterns for both processes are similar, indicating that the buy and sell counts may have a correlation structure. We can also make some observations from the two-dimensional graphs. From figure 2, the left panel is the scatter plot for the number of buy and sell transactions, which shows a positive correlation between the counts. The right panel is the scatter plot obtained by normalizing and ranking the original data, that is, pseudo-observation. Here the transformation was performed by the empirical cumulative distribution function (cdf). From this figure, we can see that the data points in the upper left and lower right corners of the

figure are relatively sparse, while the data points are concentrated on the main diagonal line Buy = Sell, which also tells us that the data is positively correlated.

These observations indicate that a PIN model with dependency is needed. The first contribution of this paper is to provide such a PIN model by implementing copula functions in the PIN model. This allows us to extend the classical PIN models to have more complicated interactions between buys and sells, which is closer to the real world as indicated by the empirical data.

This generalization, however, comes with a price that the model is much harder to fit statistically due to the existence of copula functions. Therefore, the second contribution of this paper is to provide an easy-to-implement algorithm for fitting this model. In particular, the expectation conditional maximization (ECM) method, an advanced version of the classical expectation maximum method (EM), is introduced for this statistical problem. Further numerical studies verify the capability of the proposed algorithm.

The model and the algorithm are put to the test through empirical data, which fits well for larger firms. However, for smaller firms, due to the high possibility of having no trades in a time period, both classical and proposed PIN models fail to fit the data as the Poisson distribution limits the range of probabilities for no trades. Thus, the third contribution of this paper is to introduce the zero-inflated Poisson distribution into the model, which allows the model to better capture the data when there is a considerably large proportion of no trade period in the data. The fitting capability of this zero-inflated generalization is verified through empirical data, especially for small firms. Interestingly, the empirical studies show that it is possible that a given type of information can simultaneously increase the probability of no trade as well as the number of buys and sells, which is contrary to the "good news increases buys and bad news increases sells" concept used in past literature such as [1].

The remainder of this paper is structured as follows. Section 2 presents the proposed PIN model. Section 3 discusses the corresponding estimating algorithm for the proposed model, and presents related numerical studies. Section 4 puts the model to the test through empirical studies, and discusses related observations. Section 5 concludes. Technical contents are deferred to the Appendix.

2 Proposed PIN Models

We will introduce the proposed model in three parts. First, we will briefly review the classical PIN model and copula function in Sections 2.1 and 2.2, respectively. Then, in Section 2.3, we will show how our proposed model integrates the copula function into the existing PIN model. Finally, we will introduce the zero-inflated Poisson distribution, and discuss how it is integrated into the proposed model in Section 2.4.

2.1 Review of Classical PIN Models

[1] first developed the PIN model, which is a commonly used financial model to measure information asymmetry. Figure 3 shows the structure of the PIN model in its simplest form. In short, under this model, for each trading period, the model first assumes a $1-\alpha$ probability for the market to have news; if there is news, the model further assumes a $1-\delta$ probability for it to be good news and a δ probability for it to be bad news. Finally, given there is no news, good news, or bad news in the i-th period, the model assumes that the number of buys, B_i, and the number of sells, S_i, is a joint independent Poisson random variables with parameters depending on the news scenario.

Mathematically speaking, the PIN model above essentially assumes that the pair (B, S) follows a mixture distribution with three components, each representing "bad news", "good

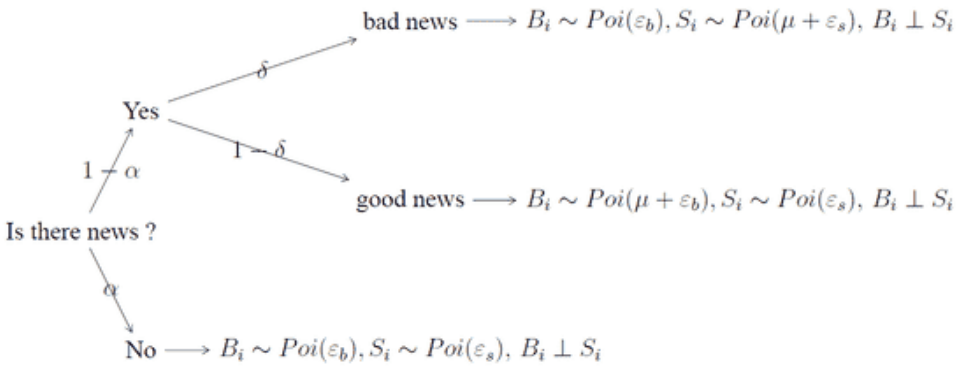


Figure 3: Structure of the PIN model in [1]

news" and "no news" market. To be more specific, the model assumes that the likelihood function of (B, S) is

$$\begin{aligned} f(B,S|\theta) = {} & \alpha e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ & + (1-\alpha)(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ & + (1-\alpha)\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^S}{S!}, \end{aligned} \tag{1}$$

where $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$. By such, given the data $(B_1, S_1), \cdots, (B_n, S_n)$, we can estimate θ by its maximum likelihood estimator (MLE)

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\Theta} \sum_{i=1}^{n} L(\theta|B_i, S_i), \tag{2}$$

which can be done through the Expectation-Maximization (EM) algorithm; see [4]. We will get back to this issue in Section 3.

It is worth noticing that more advanced PIN models have been proposed in the literature. To name a few, [2] provides a PIN model with six mixture components to handle the symmetric order flow shock; [5] develops a PIN model with time-varying parameters, and [6] proposes a data-driven PIN model with an adaptive number of mixture components. Though having different convexity, all of them assume that B and S are independent given the news scenario, so all of them face the issue discussed in the introduction. For simplicity, in this paper, we will only demonstrate the implementation of the copula function into the PIN model by [1], but all other PIN models mentioned above can be improved by a similar implementation. For other recent developments in PIN, see [7, 8] for computational researches, as well as [9, 10] for empirical studies. See also [11] for an interesting study on the PIN in COVID-19 era.

2.2 Review of Copula

One of the main purposes of this paper is to capture the dependent structure between buys and sells through copula. To do so, let us first briefly review the concept of copula and the

copula functions used in this paper. For generality, here we present copula models with p random variables; we will set $p = 2$ when applying the copulas on the joint model of buy and sell order flows.

In short, copula is a way to allow us to model marginal distribution and joint relationships separately. To be more precise, say we have p random variables $X_1, X_2, \cdots, X_p$, for which the cdf of X_i is $F_i(\cdot;\gamma_i)$ with parameter γ_i for $i = 1, 2, \cdots, p$. Now let the "copula" be a multivariate cdf C of the p-dimensional random variables $(F_1(X_1;\gamma_1), \cdots, F_p(X_p;\gamma_p))$ with parameter θ; in other words, for $\mathbf{u} = (u_1, \cdots, u_p)$,

$$C_\theta(u_1, \cdots, u_p) \coloneqq P\left\{F_1(X_1;\gamma_1) \le u_1, \cdots, F_p(X_p;\gamma_p) \le u_p\right\}.$$

Note two facts:

(a) Since $F_i(\cdot;\gamma_i)$ is the cdf of X_i, $F_i(X_i;\gamma_i) \backsim U(0,1)$, the uniform distribution which is independent to all γ_i. In other words, C_θ is no more than a multivariate cdf of $(U_1, \cdots, U_p)$ with $U_i \backsim U(0,1)$;

(b) Since F_i are monotone functions, for $\mathbf{x} = (x_1, \cdots, x_p)$,

$$\begin{aligned} P\left\{X_1 \le x_1, \cdots, X_p \le x_p\right\} &= P\left\{X_1 \le x_1, \cdots, X_p \le x_p\right\} \\ &= P\left\{F(X_1;\gamma_1) \le F(x_1;\gamma_1), \cdots, F(X_p;\gamma_p) \le F(x_p;\gamma_p)\right\} \\ &= C_\theta(F(x_1;\gamma_1), \cdots, F(X_p;\gamma_p)) \\ &\coloneqq H(\mathbf{x};\theta,\gamma) \end{aligned}$$

in which $\gamma = (\gamma_1, \cdots, \gamma_p)$. This means that the marginal distributions $F_i(\cdot;\gamma_i)$ and the copula C_θ uniquely determined the distribution of $(X_1, \cdots, X_p)$.

Combining (a) and (b), we can characterize the relationship between multiple random variables through copula "independent" to their marginal distributions.

By such, the copula provides a flexible way to construct random variables with different dependencies. Below are some common copula functions:

1. Normal copula

$$C_\theta^{(N)}(u_1, \cdots, u_p) = \Phi_\theta\left(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_p)\right), \tag{3}$$

where θ is a correlation matrix, and Φ_θ denotes the p-dimensional normal distribution with mean 0 and covariance matrix θ.

2. Gumbel copula

$$C_\theta^{(G)}(u_1, \cdots, u_p) = \exp\left[-\left(\sum_i (-\log u_i)^\theta\right)^{1/\theta}\right], \theta \in [1, \infty) \tag{4}$$

3. Clayton copula

$$C_\theta^{(C)}(u_1, \cdots, u_p) = \left[\max\left\{\sum_i (u_i^{-\theta} - 1) + 1, 0\right\}\right]^{-1/\theta}, \theta \in [-1, \infty)\backslash\{0\} \tag{5}$$

4. Independent copula

$$C^{(I)}(u_1, \cdots, u_p) = u_1 \cdots u_p \tag{6}$$

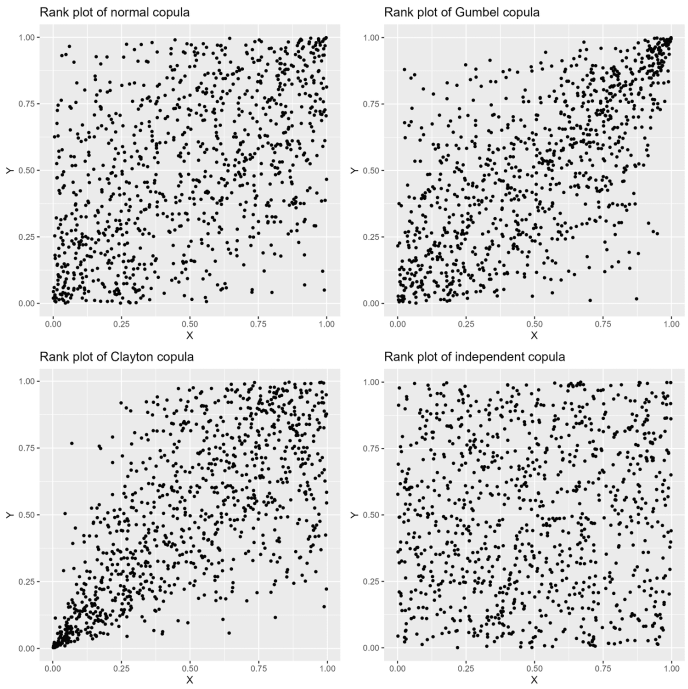


Figure 4. The rank plot of different copulas. The rank plot of a copula is the density plot when assuming all marginal distributions as uniform distribution.

The dependency corresponding to these copulas is presented in Figure 4 through rank plots. For more information on copulas, see [12] for an introduction to copulas in general, as well as [13, 14] for the implementation of copulas in financial data.

Finally, when $c_\theta(u_1, \cdots, u_p) \coloneqq \frac{\partial^p C_\theta(u_1,\cdots,u_p)}{\partial u_1 \cdots \partial u_p}$ and $f_i(x_i; \gamma_i) := \frac{\partial}{\partial x} F_i(x; \gamma_i)$ exist, the probability density function (pdf) of $(X_1, \cdots, X_p)$ is given by

$$h(\mathbf{x}; \theta, \gamma) = c_\theta(F_1(x_1; \gamma_1), \cdots, F_p(x_p; \gamma_p)) \prod_{j=1}^{p} f_j(x_j; \gamma_j). \tag{7}$$

On the other hand, when the marginals are discrete, we can express the joint probability mass function (pmf) of $(X_1, \cdots, X_p)$ as

$$h(\mathbf{x}; \theta, \gamma) = \sum_{i_1=0}^{1} \cdots \sum_{i_p=0}^{1} (-1)^{i_1+\cdots+i_p} C_\theta(F_1(x_1 - i_1; \gamma_1), \cdots, F_p(x_p - i_p; \gamma_p)). \tag{8}$$

2.3 Proposed PIN Model: The co-PIN

We are now ready to state our proposed modified PIN model, which we named "co-PIN" to indicate the use of the copula. This modification can actually be implemented in any existing PIN model; however, for easy presentation, we will focus on the modification based on the original PIN model in [1] as presented in Figure 3.

The modification is simple: For each information scenario in the original model, we change the joint distribution of (B_i, S_i) from a bivariate independent Poisson distribution to a bivariate Poisson distribution combined with a sort of copula. More precisely, for the "bad news", "good news" and "no news" scenario, we assume that:

- B_i marginally follows a Poisson distribution F_λ with parameter $\lambda = \lambda_B^-$, λ_B^+ and λ_B^0, respectively;
- S_i marginally follows a Poisson distribution F_λ with parameter $\lambda = \lambda_S^-$, λ_S^+ and λ_S^0, respectively;
- the copula of (B_i, S_i) is C_θ^-, C_θ^+ and C_θ^0, respectively.

See Figure 5 for a graphical illustration. The resulting joint distribution function is therefore

$$\begin{aligned} P\{B_i \le b, S_i \le s\} =& (1-\alpha)\delta C_\theta^- \left(F_{\lambda_B^-}(b), F_{\lambda_S^-}(s)\right) + (1-\alpha)(1-\delta) C_\theta^+ \left(F_{\lambda_B^+}(b), F_{\lambda_S^+}(s)\right) \\ &+ \alpha C_\theta^0 \left(F_{\lambda_B^0}(b), F_{\lambda_S^0}(s)\right), \end{aligned} \tag{9}$$

in which F_λ denotes the Poisson distribution with parameter λ.

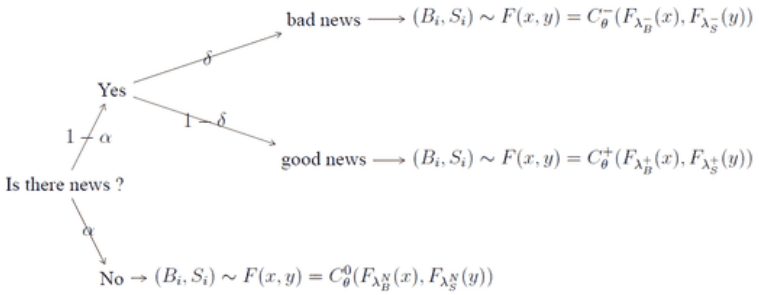


Figure 5: Structure of Proposed co-PIN Model

Note that the core concept of co-PIN is still the same as the original PIN model: the distribution of the order flows is distinct under different information scenarios. However, the original PIN model assumes that B_i and S_i are independent, or equivalently follow an independent copula, in each of the scenarios. The co-PIN, however, allows them to be dependent through a copula and even allows the copula to have different forms and parameters under different scenarios.

The proposed PIN model does require the user to choose a combination of copulas C_θ^-, C_θ^+ and C_θ^0, which means that we need a way to "compare" different copulas in order to "select" a suitable combination. As different copulas have different ranges of parameter θ, so directly comparing θ across different copula functions is not suitable. Therefore, a common way for the comparison across different copulas is through the association measure of Kendall's tau, which is defined as

$$\tau(X_1, X_2) = 4 \iint_{[0,1]^2} C(u,v)\, dC(u,v) - 1. \tag{10}$$

This measure will be used when we discuss the comparison between copulas in the later sections.

2.4 Zero-Inflated Poisson Distribution

So far, we follow the majority PIN models which assume that the order flows to be marginally Poisson distributed, under which the probability of no trade is determined by the Poisson distribution. However, as our empirical studies in Section 4 show, in some data, especially

for one of the small firms or with a short time span, the number of data points with no trade is considerably larger than the number predicted by the Poisson distribution. In other words, there exists data with considerably many no-trade data points such that a Poisson distribution is not a suitable choice.

Therefore, we introduce the zero-inflated Poisson (ZIP), which is essentially a Poisson distribution with its probability of zero being "inflated". More precisely, the pmf of a ZIP is given by

$$P\{X = x\} = \begin{cases} \phi + (1-\phi)e^{-\lambda}, & x = 0, \\ (1-\phi)\frac{\lambda^x e^{-\lambda}}{x!}, & x \geq 1, \end{cases} \tag{11}$$

where ϕ is the zero-inflated parameter. It is easy to see that ZIP is equal to a Poisson distribution with parameter λ if $\phi = 0$. In addition, for any $\phi > 0$, the ZIP will have a larger probability $P\{X = 0\}$ compared to the Poisson distribution with the same λ. For similar financial applications regarding ZIP and copulas, see [15, 16].

By such, if we find a data with considerably many no-trade data points, we can replace the Poisson distributions with ZIP. We will demonstrate the capability of this model in Section 4.

3 Estimation Algorithm

3.1 The ECM Algorithm

To apply the co-PIN to the real data, we will need a way to estimate the parameters through the data. Traditionally, the classical PIN model has been treated as a mixture model, of which the parameters are estimated through the Expectation-Maximization (EM) algorithm. See [4] for details. Theoretically, we can treat the co-PIN as a mixture model and do the parameter estimation through the EM algorithm as well. However, due to the fact that we need to estimate both of the parameters for copula and marginals simultaneously, the corresponding EM algorithm is much harder than the one for the original PIN. Therefore, we will use an advanced EM algorithm specifically designed for a mixture model with copulas. More precisely, we will use the Expectation-Condition-Maximization (ECM) algorithm proposed by [17] to fit the mixture model with copulas as in [18].

To understand the ECM algorithm, we first review how the classical EM algorithm handles mixture models. Loosely speaking, the EM algorithm is a looping algorithm that repeats an E-step and an M-step. We first have an initial guess on the parameter θ and use E-step to compute the expected log-likelihood under θ. Then we get an updated θ by maximizing the expected log-likelihood computed in the E-step, and then go back to the E-step with this updated parameter. The EM algorithm repeats this procedure until the process converges, at which it guarantees to obtain a local maximum for the likelihood function. See [19] for more details.

The ECM basically follows the same structure as the EM algorithm: compute the expectation in E-step, and then optimize the parameter in M-step. What's different for ECM is that the M-step of ECM is further divided into several CM-steps, with each CM-step only optimizing some parameters "conditioning" on the other given parameters. Note that, the M-step in the EM algorithm requires optimizing "all" parameters simultaneously, while each CM-step in the ECM algorithm only deals with "some" parameters. By such, the ECM greatly reduces the computational complexity for likelihood functions with a large number of parameters.

To be more precise, consider a mixture model with m components and pdf

$$g(\mathbf{x}; \pi, \theta, \gamma) = \sum_{j=1}^{m} \pi_j h_j(\mathbf{x}; \theta_j, \gamma_j),$$

for $\pi = (\pi_1, \cdots, \pi_m)$, $\theta = (\theta_1, \cdots, \theta_m)$ and $\gamma = (\gamma_1, \cdots, \gamma_m)$, where π_j, θ_j and γ_j are the parameters for the weight, the copula, and the marginals of the j-th component, respectively. By such, given data $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, the corresponding likelihood function is

$$L(\pi, \theta, \gamma, \pi) = \prod_{i=1}^{n} g(\mathbf{x}_i; \pi, \theta, \gamma) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{m} \pi_j h_j(\mathbf{x}_i; \theta_j, \gamma_j) \right\}. \tag{12}$$

By such, given a guess of parameters $(\pi^{(l)}, \theta^{(l)}, \gamma^{(l)})$ at the beginning of the iteration, the ECM updates the parameters through the following:

1. (E-step): Calculate $w_{ij}^{(l+1)} = \dfrac{\pi_j^{(l)} h_j(\mathbf{x}_i; \theta_j^{(l)}, \gamma_j^{(l)})}{\sum_{j=1}^{m} \pi_j^{(l)} h_j(\mathbf{x}_i; \theta_j^{(l)}, \gamma_j^{(l)})} \; (i = 1, \cdots, n, \; j = 1, \cdots, m)$

2. (CM-step1, maximize w.r.t. π): Set $\pi_j^{(l+1)} = \dfrac{1}{n} \sum_{i=1}^{n} w_{ij}^{(l+1)} \; (j = 1, \cdots, m)$

3. (CM-step2, maximize w.r.t. γ) : $\gamma^{(l+1)} = \arg\max_{\gamma} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{ij}^{(l+1)} \log h_j(\mathbf{x}_i; \theta_j^{(l)}, \gamma_j)$

4. (CM-step3, maximize w.r.t. θ) : $\theta^{(l+1)} = \arg\max_{\theta} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{ij}^{(l+1)} \log h_j(\mathbf{x}_i; \theta_j, \gamma_j^{(l+1)})$

See Appendix A for the pseudo-code of ECM.

We will therefore use this ECM algorithm to fit the parameters of our new PIN model. In this case, for (12), m would be 3, γ would be all the λ's for the Poisson distribution, and h_j is determined by (8) with F_i being the cdf of a Poisson random variable.

3.2 Simulation Studies

Before we apply empirical data, we will need to check whether the ECM algorithm performs well in the circumstances we propose. More precisely, we need to check two things: first, the ECM can help us identify the correct number of mixture components and their copula combination (as the copula types for each of the mixture components are unknown beforehand), and second, the ECM can help us estimate the parameters accurately. We will do these through two different simulation studies provided below.

3.2.1 A Case of Normal Copulas

We first show that by combining ECM and Akaike information criterion (AIC), we can identify the correct number of mixture components, as well as the correct combination of the copula across the mixture components. Here, we present one of the simulation results for a mixture model with three components of normal copulas; results for different numbers of mixture components and different copula combinations are generally the same.

For this simulation study, we first simulate data with 450 points. The data is generated from bivariate normal copula $C_{\theta}^{(N)}$ with Poisson marginal $F_{\lambda}(x)$. The first 150 points were generated from $C_{-0.4}^{(N)}(F_8(x), F_{12}(y))$, next 150 points were generated from

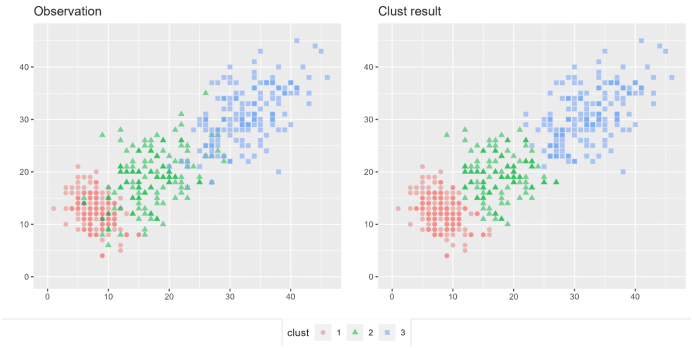


Figure 6. Scatter plot of the clustering result of the NNN data. Left panel: original data. Right panel: clustering result of model 8 (3 clusters with Poisson marginal and normal copula). The different groups are presented in different colors.

$C_{0.3}^{(N)}(F_{17}(x), F_{19}(y))$, and the last 150 points were generated from $C_{0.6}^{(N)}(F_{32}(x), F_{30}(y))$. Hereafter, we will refer to this data as NNN data. The left panel in Figure 6 shows the scatter plot of the NNN data.

After the simulation, we fit the data and do the parameter estimation through ECM under various numbers of mixture components and different copula combinations. [1] Table 1 summarizes the AIC computed when fitting various models with different copula combinations to the simulated data; for detailed results of the model fitting and parameter estimations, see Appendix B. From Table 1, we can observe that the independent models (model 7, 14, 18) have the higher AIC values when we fix the number of clusters to fit, which means that the independent model has the worst fits. Meanwhile, we also notice that no matter which copula combination was chosen, the marginal estimation result does not have much difference. Moreover, if the model and the number of clusters are correct, the marginal parameters are relatively close to the true parameters. This is also the reason why we first separate the data based on one of the marginals in the fitting procedure. From the AIC value, we conclude that model 8 is the best, and its model setting is the same as the data generation setting. This justifies the capability of ECM in conjunction with AIC to find the correct number of mixtures as well as their copula combinations.

After the parameters are estimated, we can calculate the model marginal cdf using the following model:

$$G(\mathbf{x};\psi,\pi) = \sum_{j=1}^{m} \pi_j H(\mathbf{x};\theta,\gamma) = \sum_{j=1}^{m} \pi_j C_{\theta_j}(F_{1j}(x_1;\gamma_{1j}),\cdots,F_{pj}(x_p;\gamma_{pj})) \tag{13}$$

$$\Longrightarrow G(x_i;\psi,\pi) = \sum_{j=1}^{m} \pi_j F_{ij}(x_i;\gamma_{ij}) \quad i = 1,\cdots,p. \tag{14}$$

Figure 7 shows the empirical marginal cdf (black) and the model marginal cdf (red). The two cdf lines look similar.

Model 8 correctly identifies not only the number of mixture components and their copulas but also the corresponding components for each of the data points. Figure 6 shows the true clustering of the simulated data and the clustering based on Model 8.[2] In general, the clustering result has perfectly identified the 3 clusters in the NNN data.

To check whether Model 8 also obtains the marginal distribution right, we plot the empirical marginal cdf and the model marginal cdf of each cluster. Figure 8 shows the result.

[1] For the initial value of the ECM, we use one of the marginal data, fit a one-dimensional Poisson mixture model on this marginal, and use the estimated parameters as our initial parameters for ECM.

[2] Since cluster 2 has more overlap with other clusters, it has more points with incorrect labels.

Panel A. Models with Two Mixture Components			
	normal	Gumbel	Clayton
normal	(1) 6009	(2) 6022	(3) 6058
Gumbel		(4) 6012	(5) 6037
Clayton			(6) 6084
All Independent: (7) 6189			
Panel B. Models with Three Mixture Components in which the first component is normal copula			
	normal	Gumbel	Clayton
normal	**(8) 5855**	(9) 5858	(10) 5868
Gumbel		(11) 5858	(12) 5869
Clayton			(13) 5869
All Independent: (14) 5916			
Panel C. Models with Four Mixture Components in which the first two components are normal copulas			
	normal	Gumbel	Clayton
normal	(15) 5863	(16) 5861	(17) 5859
All Independent: (18) 5922			

Table 1: The fitting result for the simulation study

This table summarizes the AIC computed when fitting various models to the simulated data in Section 3.2.1. The number in the brackets are the model label, following by the computed AIC of that model. Each panel corresponds to models with different numbers of mixture components, while the rows and columns indicate the copula for the last two components. For example, Model #12 in Panel B corresponds to the model with three mixture components, of which the copulas are normal/Gumbel/Clayton respectively. For comparison, we also present the AIC corresponds to the models with all components being independent copulas. See Table 4 in Appendix B for more details.

Cluster 3 fits well. Note that there are more points of mislabeling in cluster 2, so the two cdf lines have a slight difference.

To sum up, in this case study, the ECM in conjunction with AIC can identify the number of mixture components, the corresponding copulas, the corresponding mixture components for each data point, and the marginal distributions reasonably well.

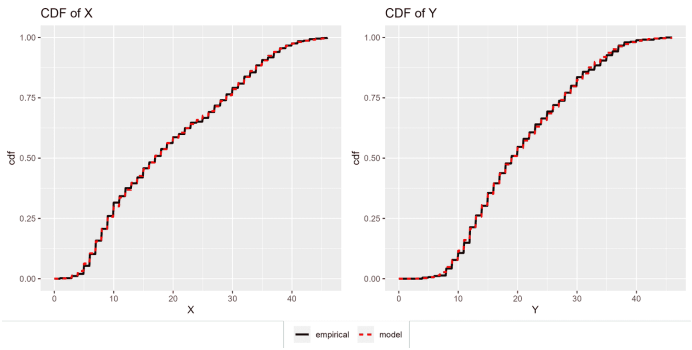


Figure 7. Comparison of the marginal cdf in NNN data. Black real line: empirical marginal cdf. Red dash line: mixture marginal cdf of model 8. The empirical marginal cdf line and the model marginal cdf line look similar.

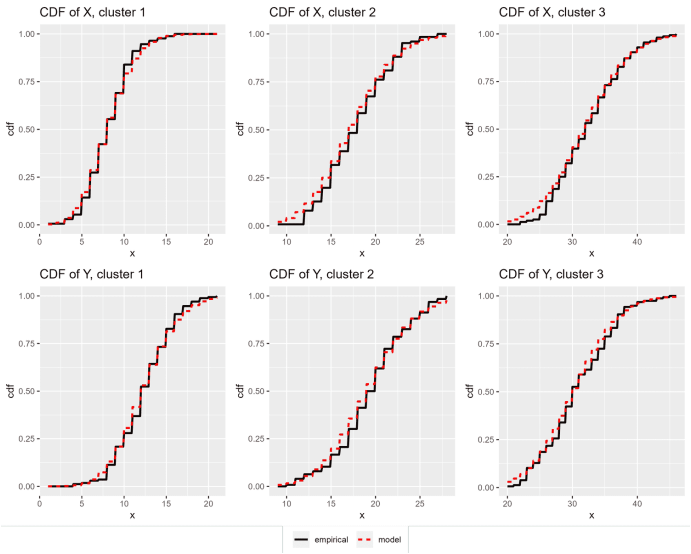


Figure 8. cdf of the NNN data in each group. Black real line: empirical marginal cdf. Red dash line: marginal cdf of the model 8.

3.2.2 Large scale experiment

To verify that the fitting method is reasonable in general, not just for a particular setting, we repeatedly generate 1000 datasets with 500 two-dimensional data points with 2 clusters and then fit the models. We construct the data from specified marginals (Poisson) and copulas (normal, Gumbel, Clayton), then fit the simulation data using the procedure introduced in Section 3.2.1 with the same model setting. Finally, we use histogram to check the difference. Here we set the marginal parameter in the first cluster from $\lambda_{1B} \sim U(25, 75), \lambda_{1S} \sim U(25, 75)$, and marginal parameters in second cluster from $\lambda_{2B} \sim \lambda_{1B} + U(-25, 25), \lambda_{2S} \sim \lambda_{2S} + U(-25, 25)$. The ranges for the normal copula is $\theta \in (-0.8, 0.8)$, and for the Gumbel and Clayton copula are $\theta \in (1.1, 10)$. The range of weight is $w \in (0.1, 0.9)$.

Figure 9 shows the histogram of the absolute difference of the parameters. We can observe that the differences between the estimated parameters and the true parameters are small. A few differences are relatively large due to very extreme combinations of copula parameters and weights.

Figure 10, on the other hand, shows a scatter plot of error rates and distances between two clusters in 1000 simulated datasets. Here the distances were defined by the parameters Euclidean distance between two clusters. It can be observed that as the distance increases, the classification performance improves. And even when the distance is relatively small, the classification error rate is only about 30%. This is due to the distinct structure of the copula, which provides additional information for clustering. Based on the simulation results, the copula-based clustering via ECM algorithm is an effective method.

4 Empirical Studies

For the empirical data, we use the daily trading records of AAPL (large) and NTE (small) stocks in 2008, which were collected every 5 minutes. There are 249 days in total. Since there may be odd-lot trades outside of regular trading hours, we only use data within U.S. stock market trading hours (9:30 a.m. to 4:00 p.m.), resulting in about 78 data points per day. We will implement the proposed model in Section 2 on these two models, both with Poisson

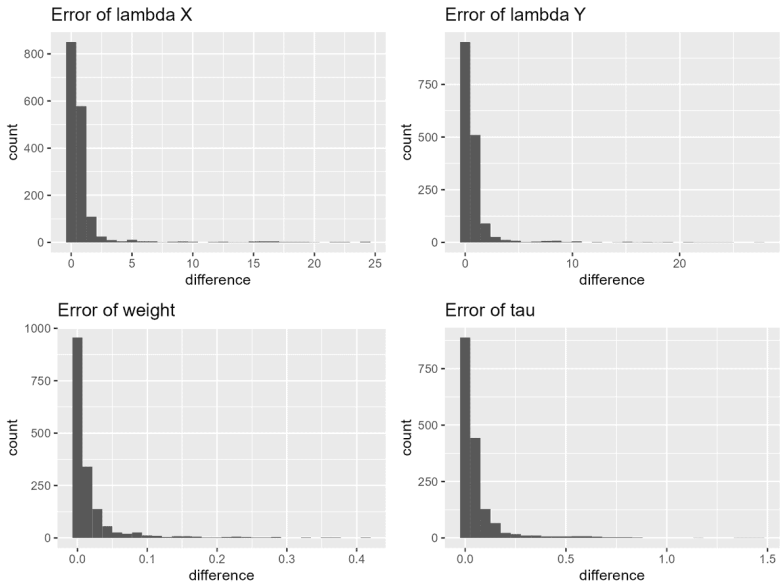


Figure 9: Histogram of absolute error of parameters. The upper-left and upper-right show the marginal absolute error. The bottom-left shows the weight absolute error. The bottom-right shows the absolute error of the Kendall's tau defined in (10).

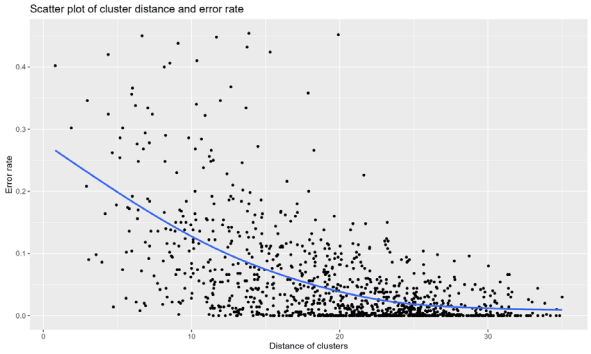


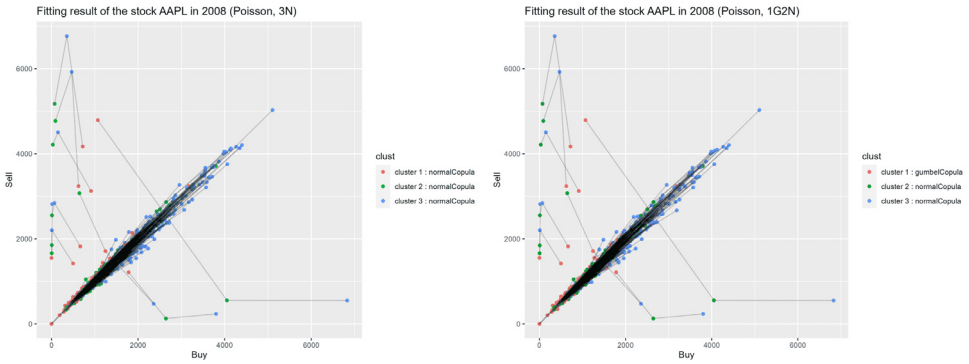
Figure 10. Scatter plot of classification error rate vs. cluster distance. Each point represents a dataset and the corresponding fitted result. The blue line is the generalized additive model (GAM) fit to these points.

marginals as well as ZIP marginals proposed in Section 2.4. The parameter estimations, on the other hand, are all executed through the ECM algorithm in Section 3.

4.1 Stock AAPL in 2008

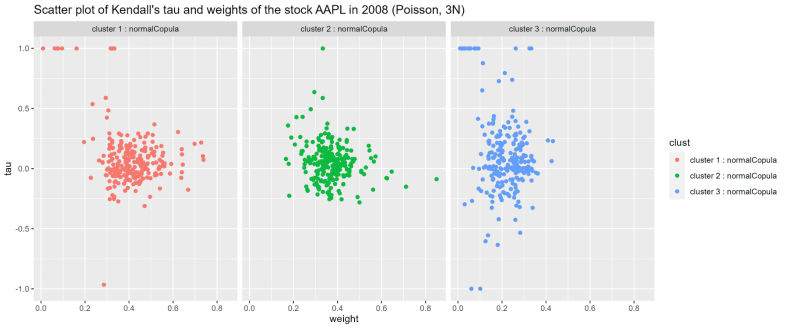
4.1.1 Poisson marginal

First, we fit 3 clusters with Poisson marginal distribution and normal copulas. Figure 11a shows the fitting result of the stock AAPL in 2008. Each point represents the (λ_B, λ_S) of each cluster, and all lines connect 3 points of the fit in one day. We noticed that the data points are mostly concentrated on the 45-degree line. Intuitively, this suggests that when there is no news in the market, the trading volume tends to be relatively low. And when there is news in the market, whether it's good news or bad news, trading volume tends to increase.

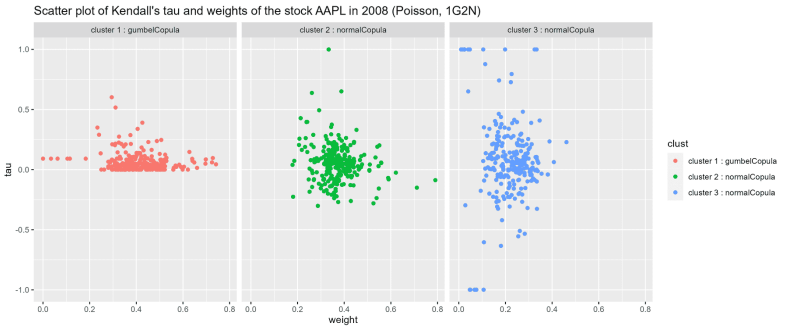


(a) 3 Normal Copulas (b) 1 Gumbel + 2 Normal Copulas

Figure 11: Poisson rate of the AAPL stock in 2008 under different coupla combinations. All lines connect 3 points of the fit in one day.



(a) 3 Normal Copulas



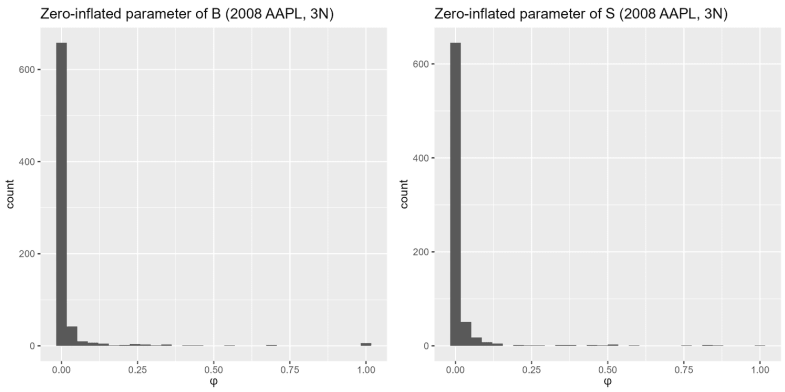
(b) 1 Gumbel + 2 Normal Copulas

Figure 12: Scatter plot of Kendall's tau and the weights of the AAPL stock in 2008 under different copula combinations. 3 clusters have a similar copula structure.

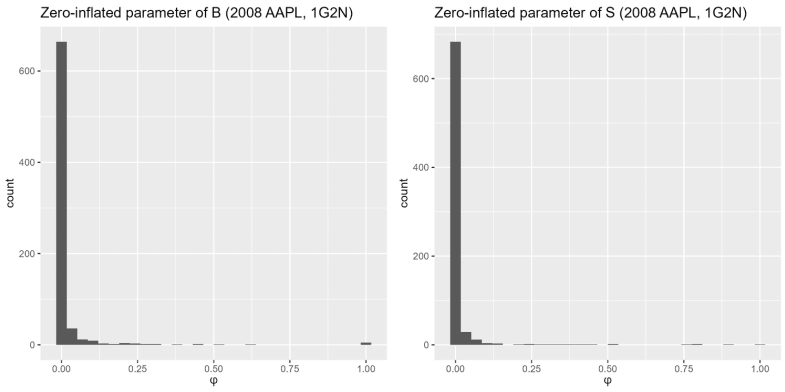
Next, we observe the relation between Kendall's tau and the weights of each cluster from the scatter plot (Figure 12a). We can see that the weights are not very different between the 3 clusters. Most of the Kendall's tau values are between 0.5 and -0.5. Cluster 3 has a higher correlation than the other clusters, but the weights are relatively low. In conclusion, the 3 clusters seem to have a similar copula structure.

Now we change the combination of copula settings, fitting 3 clusters with Poisson marginal distribution and 1 Gumbel copula, 2 normal copulas. Figure 11b shows the Poisson rate of the fit. Compared to Figure 11a, we can see that there is not much difference in the marginal parameters. Changing the copula combination does not affect the marginal parameters.

Figure 12b shows the scatter plot of Kendall's tau and the weights. We can observe that for cluster 2 and cluster 3, Kendall's tau values are similar to the 3 normal copulas setting. However, Kendall's tau values in cluster 1 are all positive, since Gumbel copula can only describe the positive correlation.



(a) 3 Normal Copulas



(b) 1 Gumbel + 2 Normal Copulas

Figure 13: Histogram of zero-inflated parameters of the AAPL stock in 2008

4.1.2 Zero-inflated Poisson marginal

We then repeat the procedure with Poisson marginals replaced by zero-inflated Poisson marginals. The resulting Poisson rate and Kendall's tau are almost identical to the case with Poisson marginals in Figures 11 and 12, and therefore are omitted.

The results on the zero-inflated parameter ϕ in (11), on the other hand, are presented in Figure 13. Figure 13a shows the histogram of zero-inflated parameters (3 clusters with normal copula), we observed that most of the zero-inflated parameters are zero. And Figure 13b shows the histogram of zero-inflated parameters (3 clusters with 1 Gumbel copula, 2 normal copula). We also observed that most of the zero-inflated parameters are 0, different combinations of copula do not affect the zero-inflated parameters.

To compare the Poisson and zero-inflated Poisson, we select one day (2008/05/29) of the stock AAPL and then check the estimated parameters, which were shown in Table 2. We can see that the mean numbers of buys (governed by λ_B) and sells (governed by λ_S) are the almost same in three mixture components, and the zero-inflated parameters ϕ_B and ϕ_S are both close to 0. Hence, zero-inflated Poisson didn't improve the fitting for large firms. This can also be verified by the AIC values.

Marginal	Copula	λ_B	λ_S	θ_j	τ_j	π_j	AIC
	normal	475.4278	490.2999	0.0908	0.0579	0.3889	
Poisson	normal	766.1322	790.2893	−0.0231	−0.0147	0.3761	5450.984
	normal	1310.7000	1172.6000	0.5074	0.3388	0.2351	
	normal	472.1086	491.2782	0.2196	0.1409	0.4395	
ZIP	normal	782.1377	790.0577	−0.0580	−0.0369	0.4150	5526.589
	normal	1310.0882	1171.5819	0.5480	0.3692	0.1455	
		ϕ_B	ϕ_S				
	normal	0.0169	0.0739				
ZIP	normal	0.0000	0.0032				
	normal	0.0000	0.0000				

Table 2: The fitting result of the stock AAPL (2008/05/29)

4.2 Stock NTE in 2008

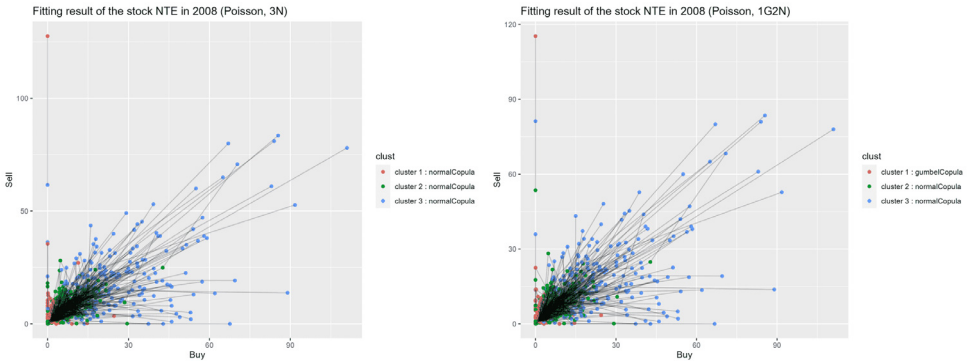
So far, we have presented the capability of our model and algorithm through the stock AAPL, which is a stock with relatively high liquidity. However, not all stocks have such large trading volumes. Hence, we use stock NTE to see if the conclusion is the same for small firms.

4.2.1 Poisson marginal

Figure 14a shows the Poisson rates of the stock NTE in 2008, which was fitted with 3 clusters of Poisson marginal distribution and normal copulas. We can also see that the data points are mostly concentrated on the 45-degree line. However, since the transaction counts are relatively small, most of the estimated rates are lower than 10.

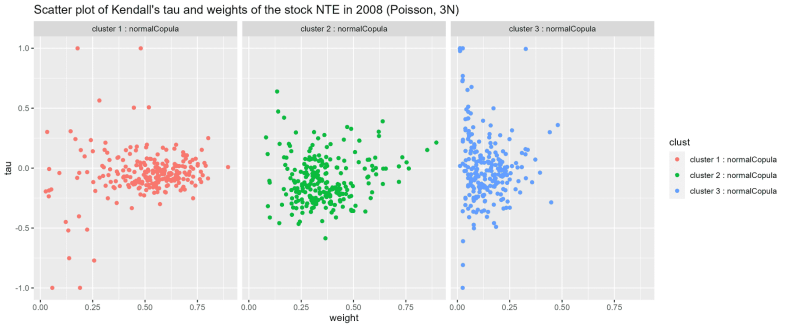
Next, we observe the relation between Kendall's tau and the weights of each cluster from the scatter plot (Figure 15a). We can also see that the weights are not very different between the 3 clusters. However, for the cluster with higher transaction rates (cluster 3), the weights are lower than the other clusters, and Kendall's tau value was greater than 0.5 or less than -0.5, concluding that the data have a stronger correlation.

For the Poisson rates fit with 1 Gumbel copula and 2 normal copulas, the results are shown in Figures 14b and 15b. Note the results are highly similar to the case with 3 normal copulas.

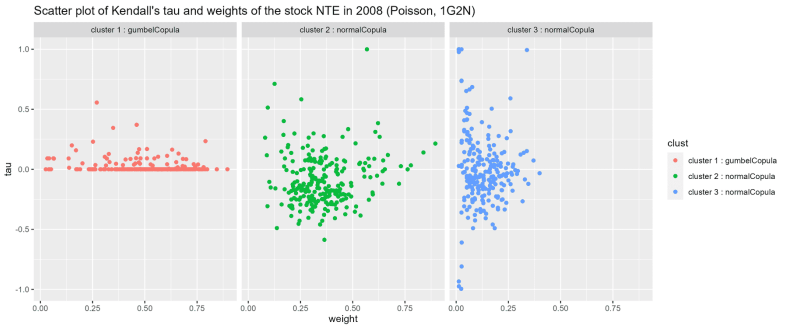


(a) 3 Normal Copulas (b) 1 Gumbel + 2 Normal Copulas

Figure 14: Poisson rate of the NTE stock in 2008 under different copula combinations. All lines connect 3 points of the fit in one day.

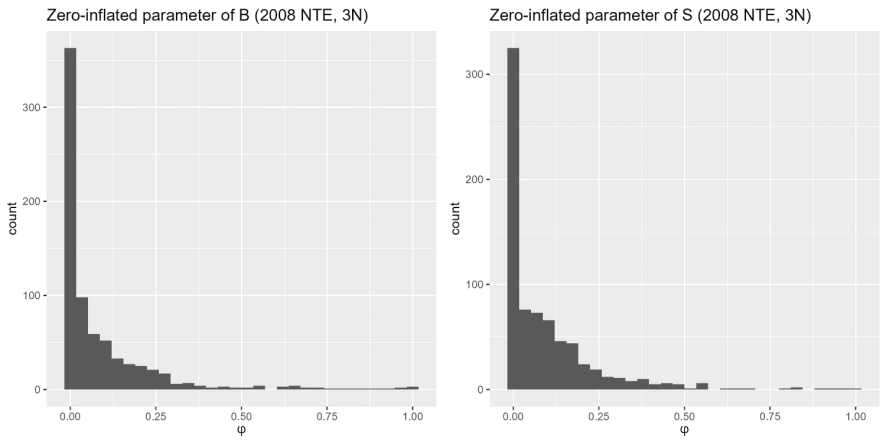


(a) Poisson marginal

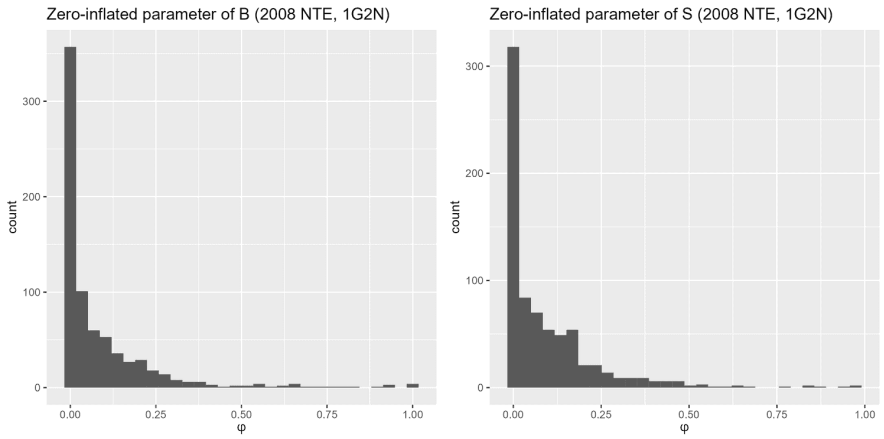


(b) Poisson marginal

Figure 15: Scatter plot of Kendall's tau and the weights of the NTE stock in 2008 under different copula combinations.



(a) 3 normal



(b) 1 Gumbel, 2 normal

Figure 16: Histogram of zero-inflated parameters of the NTE stock in 2008.

4.2.2 Zero-inflated Poisson marginal

As discussed in Section 2.4, the ill-fit of the model with Poisson marginals is mainly due to the fact of a large amount of no trades in the data. In order to overcome the fact that most of the data are concentrated at 0, leading to under-estimate the transaction rates, we fit the stock NTE in 2008 with zero-inflated Poisson marginal introduced in Section 2.4 instead. Similar to the case of AAPL in Section 4.1.2, the results in Poisson rate and Kendell's tau are highly similar to the Poisson marginal cases presented in Figures 14 and 15, and therefore are omitted. However, for the zero-inflated parameter ϕ in (11), Figure 16 shows that there are some zero-inflated parameters greater than 0.1, proving that there are more zeros in the data than expected by the models with Poisson marginals. To demonstrate that the zero-inflated Poisson has a better fit, we select one day (2008/05/29) of the stock NTE and then check the cluster result and the marginal cdf. The estimated parameters are shown in Table 3.

Marginal	Copula	λ_B	λ_S	θ_j	τ_j	π_j	AIC
	normal	1.9244	1.6620	−0.1563	−0.0999	0.5250	
Poisson	normal	5.9193	12.3169	0.0196	0.0125	0.2777	925.6160
	normal	12.8452	4.2603	−0.0846	−0.0540	0.1973	
	normal	2.5152	3.4065	0.0623	0.0397	0.5315	
ZIP	normal	7.9531	13.7816	−0.2545	−0.1638	0.2346	883.4752
	normal	12.8518	3.7130	−0.1769	−0.1132	0.2339	
		ϕ_B	ϕ_S				
	normal	0.0607	0.3849				
ZIP	normal	0.0690	0.0010				
	normal	0.2777	0.0353				

Table 3: The fitting result of the stock NTE (2008/05/29)

Figure 17 shows the different clustering results. The left panel is the result of the Poisson marginal, and the right panel is the result of the zero-inflated Poisson marginal. We can see that for zero-inflated Poisson marginal, some zeros have been clustered into cluster 3, indicating that some traders tend not to trade when the market has news.

It is worth noticing that in Table 3, the third normal component has the highest λ_B, but also the highest ϕ_B among all three mixture components. This means that, for the good news state, the probability of no buy (governed by ϕ_B) increases, but given there is a buy, the mean number of buys (governed by λ_B) also increases. On the other hand, the second normal component has the highest λ_S, but not the lowest ϕ_S. In other words, for the bad news state, the probability of no sell is the smallest, and the mean number of sells is the largest. This observation is different from the common belief that [1] and others have assumed, and the asymmetric phenomenon might have some relationship with the perspective theory, under which people have asymmetric behavior under good and bad news.

(a) Poisson marginal

(b) Zero-inflated Poisson marginal

Figure 17: Clustering result of the stock NTE in 2008/05/29 (3 normal)

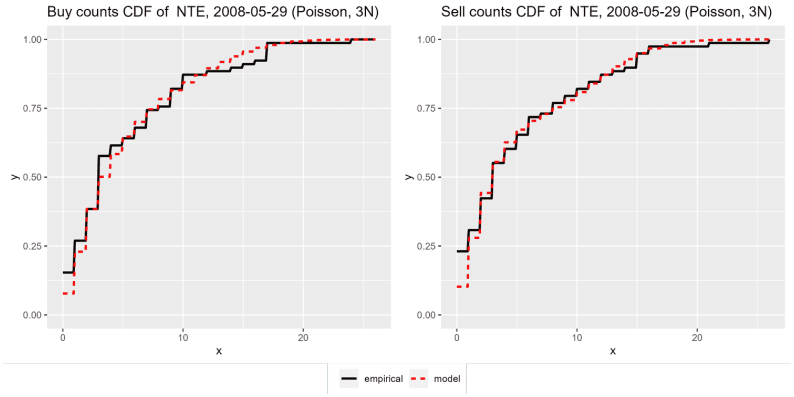


Figure 18: Marginal cdf of the NTE stock in 2008/05/29 (Poisson, 3 normal)

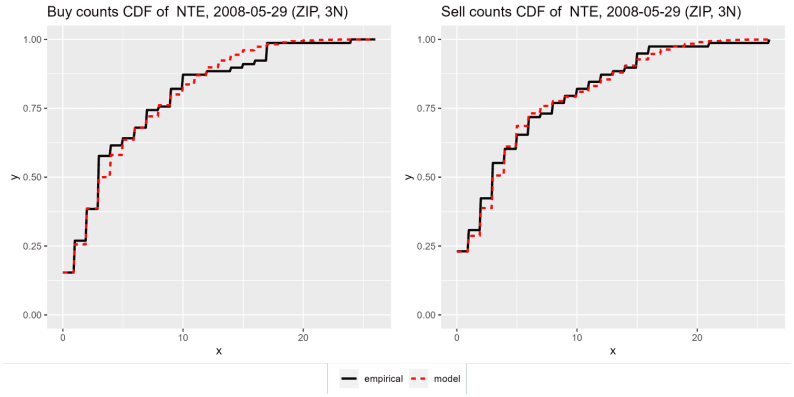


Figure 19: Marginal cdf of the NTE stock in 2008/05/29 (ZIP, 3 normal)

Finally, we check the marginal cdf. Figure 18 shows the mixture marginal cdf of the Poisson. We observe that the fitted and empirical cdfs have a relatively large difference near $x = 0$, which is because there are more zeros in the data than expected by a Poisson distribution. And we observe in Figure 19 that the mixture marginal cdf of zero-inflated Poisson overcomes this bias.

We can also check the marginal cdf in each cluster. Figure 20 shows the marginal cdf of the Poisson marginal in each cluster. We find that the marginal fits have a difference close to 0 with the empirical cdf in cluster 1. And for clusters 2 and 3, there are no zeros in these clusters. However, in Figure 21, which shows the marginal cdf of the zero-inflated Poisson in each cluster, the fitted cdf appears to have a better fit than the Poisson marginal. In addition, there are some zeros in cluster 3, which also tells us that some traders tend not to trade even when there is news in the market.

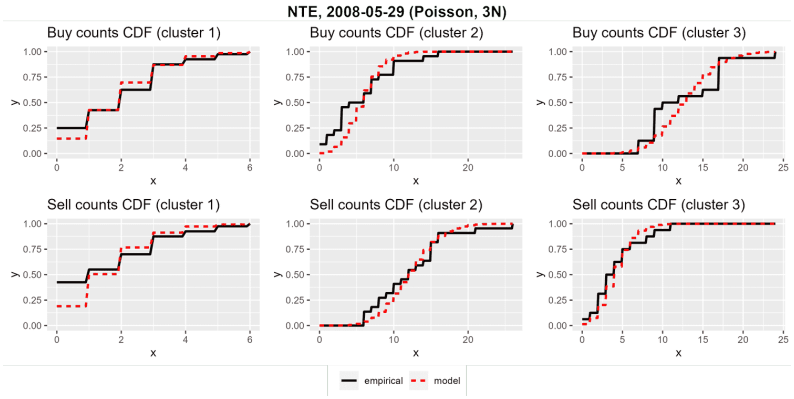


Figure 20: Marginal cdf of each cluster of the NTE stock in 2008/05/29 (Poisson, 3 normal)

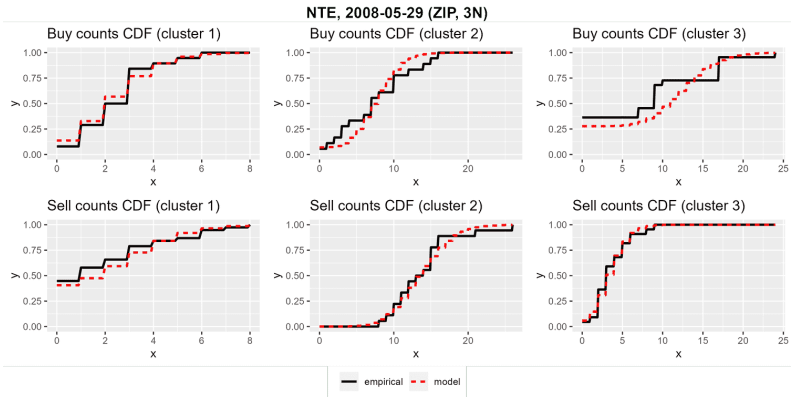


Figure 21: Marginal cdf of each cluster of the NTE stock in 2008/05/29 (ZIP, 3 normal)

5 Conclusion

In our paper, we have proposed a method to evaluate the PIN model with the correlation structure by using the copula function and the ECM algorithm. From the simulation result in Section 3.2, we can conclude that the ECM algorithm is effective and reliable. From the real data analysis, we observed that there is a correlation in the PIN model, proving that the independent assumption was too strict. Moreover, no matter what copula combination is used, the marginal fitting result does not have much difference.

From Section 4.1 we observe that for a large company, no matter which marginal distribution to use, the fitting result is not much different. Therefore, we only need to determine which combination of copula functions to use. Another observation from Section 4.2 is that as the frequency of counts increases, the correlation also tends to increase. Note that for the stocks with relatively small counts, negative correlation may occur, making the selection of an appropriate copula model an important issue.

In future work, we aim to modify the code to fit the rotate copula and find a suitable method to obtain initial parameters. From the empirical studies results, the issue of misspecification in the mixture copula model is also an interesting topic for further research.

References

[1] D. Easley, N.M. Kiefer, M. O'hara, J.B. Paperman, The Journal of Finance **51**, 1405 (1996)
[2] J. Duarte, L. Young, Journal of Financial Economics **91**, 119 (2009)
[3] Q. Gan, W.C. Wei, D. Johnstone, Financial Review **52**, 5 (2017)
[4] E. Lin, C.F. Lee, in *Handbook of Financial Econometrics and Statistics* (Springer, 2015), pp. 2601–2619
[5] D. Easley, R.F. Engle, M. O'Hara, L. Wu, Journal of Financial Econometrics **6**, 171 (2008)
[6] E. Lin, C.L.M. Kao, N.S. Adityarini, Review of Quantitative Finance and Accounting **57**, 411 (2021)
[7] Q. Gan, W.C. Wei, D. Johnstone, Quantitative Finance **15**, 1805 (2015)
[8] M. Ghachem, O. Ersan, The R Journal **15**, 145 (2023)
[9] A. Bird, S.A. Karolyi, T.G. Ruchti, P. Truong, Review of Finance **25**, 745 (2021)
[10] A. Frino, R. Palumbo, P. Rosati, Accounting & Finance **63**, 2597 (2023)
[11] C.O. Cepoi, V. Dragotă, R. Trifan, A. Iordache, Financial Innovation **9**, 34 (2023)
[12] R.B. Nelsen, *An introduction to copulas* (Springer science & business media, 2007)
[13] L. Hu, Applied financial economics **16**, 717 (2006)
[14] A. Di Clemente, C. Romano, Frontiers in Applied Mathematics and Statistics **7**, 642210 (2021)
[15] M. Alqawba, D. Fernando, N. Diawara, Computation **9**, 108 (2021)
[16] L. Bermúdez, D. Karlis, TEST **31**, 1082 (2022)
[17] X.L. Meng, D.B. Rubin, Biometrika **80**, 267 (1993)
[18] I. Kosmidis, D. Karlis, Statistics and computing **26**, 1079 (2016)
[19] A.P. Dempster, N.M. Laird, D.B. Rubin, Journal of the royal statistical society: series B (methodological) **39**, 1 (1977)

Appendix

A Pseudo-code for ECM Algorithm

The ECM algorithm is presented in Section 3. For readers interested in coding this algorithm, please see the following pseudo-code.

Algorithm 1 Pseudo-code for copula-based clustering via ECM algorithm

function coECM($\mathbf{x}, \gamma, \theta, \pi, \varepsilon$)
  $\gamma_j^{(0)} \leftarrow \gamma_j$; $\theta_j^{(0)} \leftarrow \theta_j$; $\psi_j^{(0)} \leftarrow (\gamma_j^{(0)}, \theta_j^{(0)})$; $\pi_j^{(0)} \leftarrow \pi_j$; $l \leftarrow 0$ ▷ Initial parameters
  while True **do**
    $w_{ij}^{(l+1)} \leftarrow \dfrac{\pi_j^{(l)} h_j(\mathbf{x}_i; \psi_j^{(l)})}{\sum_{j=1}^{m} \pi_j^{(l)} h_j(\mathbf{x}_i; \psi_j^{(l)})} \quad (i = 1, \cdots, n,\ j = 1, \cdots, m)$ ▷ E-Step
    $\pi_j^{(l+1)} \leftarrow \sum_{i=1}^{n} w_{ij}^{(l+1)} / n$ ▷ CM-Step1: update weight
    for $j = 1$ to m **do** ▷ CM-Step2: update marginal parameters
      $\gamma_j^{(l+1)} \leftarrow \arg\max_{\Gamma_j} \sum_{i=1}^{n} w_{ij}^{(l+1)} \log h_j(\mathbf{x}_i; \theta_j^{(l)}, \gamma_j)$
    end for
    for $j = 1$ to m **do** ▷ CM-Step3: update copula parameters
      $\theta_j^{(l+1)} \leftarrow \arg\max_{\Theta_j} \sum_{i=1}^{n} w_{ij}^{(l+1)} \log h_j(\mathbf{x}_i; \theta_j, \gamma_j^{(l+1)})$
    end for
    $\phi_j^{(l+1)} \leftarrow (\gamma_j^{(l+1)}, \theta_j^{(l+1)}, \pi_j^{(l+1)}) \quad (j = 1, \cdots, m)$
    if $\|\phi^{(l+1)} - \phi^{(l)}\| < \varepsilon$ **then break**
    end if
    $\psi_j^{(l+1)} \leftarrow (\gamma_j^{(l+1)}, \theta_j^{(l+1)})$; $l \leftarrow l + 1$
  end while
end function

B Detailed Results for Simulation Studies

Table 4 presents the detailed results of model fitting in Section 3.2.1, in which Model 1 through model 7 are fitted with 2 clusters, model 8 through model 14 are fitted with 3 clusters, and model 15 through model 18 are fitted with 4 clusters.

Model	Copula	λ_x	λ_y	θ_j	τ_j	π_j	AIC
True	normal	8	12	−0.4	−0.2619	0.333	-
	normal	17	19	0.3	0.1939	0.333	
	normal	32	30	0.6	0.4096	0.333	
1	normal	10.1223	14.0114	0.1879	0.1203	0.5156	6008.637
	normal	29.0878	27.9923	0.6242	0.4291	0.4844	
2	normal	10.1830	14.0045	0.1893	0.1213	0.5188	6022.347
	Gumbel	29.1556	27.8980	1.6590	0.3972	0.4812	
3	normal	10.5814	14.4619	0.2559	0.1648	0.5469	6057.732
	Clayton	30.3954	28.8821	0.9658	0.3257	0.4531	
4	Gumbel	10.1492	14.0151	1.1550	0.1342	0.5245	6012.114
	Gumbel	29.2956	27.9583	1.6698	0.4011	0.4755	
5	Gumbel	10.3507	14.4271	1.2114	0.1745	0.5474	6037.019
	Clayton	30.5241	28.7020	1.0485	0.3439	0.4526	
6	Clayton	10.1273	14.0352	0.0187	0.0092	0.5165	6084.490
	Clayton	29.6483	28.6612	0.8515	0.2986	0.4835	
7	indep.	10.5578	14.1797	NA	0	0.5420	6189.083
	indep.	29.6430	28.5575	NA	0	0.4580	
8	normal	8.2238	12.4010	−0.3765	−0.2457	0.3641	5855.479
	normal	17.3862	19.2626	0.1838	0.1177	0.2767	
	normal	32.0187	30.4074	0.5985	0.4085	0.3592	
9	normal	8.2271	12.4065	−0.3745	−0.2443	0.3647	5858.008
	normal	17.5481	19.3168	0.1972	0.1264	0.2836	
	Gumbel	32.2825	30.6532	1.6180	0.3819	0.3517	
10	normal	8.2465	12.4083	−0.3700	−0.2413	0.3664	5868.304
	normal	17.5760	19.4678	0.1913	0.1225	0.2798	
	Clayton	32.2950	30.4486	1.0951	0.3538	0.3539	
11	normal	8.2580	12.4140	−0.3672	−0.2393	0.3683	5858.247
	Gumbel	17.7265	19.4556	1.1354	0.1193	0.2821	
	Gumbel	32.3173	30.6598	1.6162	0.3813	0.3496	
12	normal	8.2787	12.4144	−0.3627	−0.2363	0.3699	5868.690
	Gumbel	17.6962	19.6300	1.1259	0.1118	0.2780	
	Clayton	32.3474	30.4274	1.0899	0.3527	0.3522	
13	normal	8.3976	12.4765	−0.3092	−0.2001	0.3783	5868.935
	Clayton	17.7777	19.6418	0.1634	0.0755	0.2679	
	Clayton	32.2777	30.4718	1.1023	0.3553	0.3538	
14	indep.	8.6049	12.6481	NA	0	0.3972	5916.117
	indep.	18.6791	20.0633	NA	0	0.2747	
	indep.	32.7723	31.1807	NA	0	0.3280	
15	normal	8.2235	12.4009	−0.3766	−0.2458	0.3641	5863.480
	normal	17.3780	19.2650	0.1821	0.1165	0.1384	
	normal	32.0171	30.4076	0.5985	0.4085	0.3592	
	normal	17.3941	19.2575	0.1857	0.1189	0.1383	
16	normal	8.2169	12.3859	−0.3807	−0.2486	0.3620	5861.210
	normal	16.4980	19.6859	0.2162	0.1387	0.2107	
	normal	32.2636	30.7156	0.5801	0.3940	0.3429	
	Gumbel	21.1989	18.9677	1.9877	0.0843	0.0843	
17	normal	8.2111	12.3842	−0.3834	−0.2505	0.3612	5859.368
	normal	16.9891	19.3886	0.2471	0.1590	0.2637	
	normal	32.0974	30.7751	0.6289	0.4330	0.3420	
	Clayton	26.4356	19.8333	4.4410	0.6895	0.0331	
18	indep.	8.6048	12.6483	NA	0	0.3972	5922.117
	indep.	18.6793	20.0634	NA	0	0.1374	
	indep.	32.7718	31.1811	NA	0	0.3280	
	indep.	18.6793	20.0634	NA	0	0.1374	

Table 4: The fitting results of NNN data