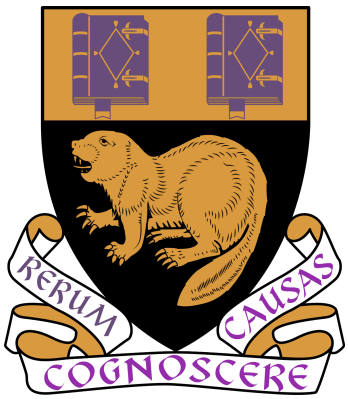
Point Processes and Integer-valued time series



Zezhun Chen

The Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

October 2023

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 4 to Chapter 8 are modified versions of published papers, and most of them are joint work with my supervisors: Prof Angelos Dassios and Dr George Tzougas. The detail information of these papers are listed in the following 'Accompanying Papers' Chapter.

Acknowledgements

Foremost, I extend my profound gratitude to my supervisors, Prof. Angelos Dassios and Dr. George Tzougas, for their remarkable patience, unwavering encouragement, and robust guidance. Prof. Dassios graciously ushered me into the realm of research and unveiled the fascinating topic of cluster point processes. His mentorship has been pivotal in honing my research skills and crafting my scientific communication and judgment. On the other hand, Dr. Tzougas introduced me to a novel integer-valued time series model, enhancing my proficiency in applying statistical models to insurance data. This work would not have materialized without their dedicated involvement and persistent advice.

I express heartfelt thanks to the London School of Economics, not only for fostering a stimulating academic environment but also for being a steadfast employer. The support processes were impeccable, from the IT-support desk at the library and general assistance from Penny Montague to teaching arrangement support from Imelda Noble and boundless academic aid from our esteemed faculty within our delightful department. They even offered free courses on various topics from other universities. Few scientists have the privilege to work under such professional and well-organized conditions. Additionally, I sincerely acknowledge the Department of Statistics for providing the studentship that enabled me to pursue my Ph.D. degree without financial concerns.

Last but not least, I extend my deepest appreciation to my parents, Shaobo Chen and Jinge Liang, for their unconditional love, support, and ongoing encouragement. My gratitude also goes to my Ph.D. cohorts: Shakeel Gavioli-Akilagun, Eduardo Ferioli-Gomes, and Wenyu Cheng, who have enriched the time we spent together in the office. Special words of gratitude are reserved for my incredible friends, with whom I played badminton and improved my physical fitness over the years: Jes Yang, William Meng, Pengxuan Zhang, Dongxuan Pang, Tony Xu, Jiacheng Tian, and many more. Thank you all for your constant presence and multifaceted inspiration.

My sincere thanks also extend to all my family members, dear friends, and everyone who has accompanied me on this unique journey.

Abstract

This thesis explores two primary themes across five scientific papers: Integer-value time series and their relationship with classical point processes.

The first part of the thesis focuses on the development and application of Integer-valued autoregressive (INAR) models, extending from univariate to multivariate cases, with applications in financial and insurance count data. In Paper A, we introduce a new family of binomial-mixed Poisson INAR model of order one, INAR(1), by incorporating a mixed Poisson component to the innovation of the classical Poisson INAR(1). This allows for the capture of overdispersion and serial correlation evident in financial count data. Furthermore, we explore its distributional properties, estimation procedure and asymptotic properties and apply the model to iceberg count data from financial system. In Paper B, we extending beyond univariate case, introducing a novel family of multivariate mixed Poisson-Generalized Inverse Gaussian INAR(1), MMPGIG-INAR(1), regression models for modelling multivariate count time series. This family of models can accommodate a wide range of dispersion and cross-sectional correlation structures due to the flexibility in the parameter setting of the Generalized Inverse Gaussian. We then illustrate different members of the MMPGIG-INAR(1) through applying the model to Local Government Property Insurance Fund data from the state of Wisconsin. In Paper C, we develop novel Expectation-Maximization estimation algorithm for maximum likelihood estimation of bivariate mixed Poisson INAR(1) model. This method is readily extensible to the multivariate case. We examine three different mixing densities, univariate gamma, bivariate Lognormal and bivariate copula and demonstrate the algorithm through fitting the same used in Paper B.

The second part of the thesis shifts focus to integer-valued approximation of classical point processes and applications of point process on covid data modelling. In Paper D, we represent the Cox process and the dynamic contagion process, which is a Hawkes process whose immigration part is a Cox process, as limit of time-series based point processes, namely integer-valued moving average model (INMA)

and Integer-valued Autoregressive Moving Average model (INARMA). This would potentially facilitate the statistical inference of classical point processes. In Paper E, we propose a new type of univariate and bivariate Integer-valued autoregressive model of order one, INAR(1), to approximate univariate and bivariate linear birth and death process with constant rates. Due to the simplicity of Markov structure of INAR model, we demonstrate through simulation study that the parameters of linear birth and death process can be estimated through Quasi-likelihood function of INAR model.

Contents

1 Introduction **10**

1.1 Integer-Valued Time series 10

1.2 Cluster Point Processes 13

1.3 Birth and Death Processes 13

2 Integer-value time series **16**

2.1 Distributional and Statistical Properties 17

2.2 Parameter estimation and Forecasting 19

2.3 INAR model with correlated innovations (Paper A) 20

2.4 Multivariate INAR model with GIG family (Paper B) 23

2.5 EM algorithm for Multivariate INAR model (Paper C) 25

3 Point Processes **31**

3.1 Cluster Point Process 33

3.2 Birth and Death Process 36

3.3 Integer-valued approximation of point processes (Paper D,E) 38

4 Paper A. A First Order Binomial Mixed Poisson Integer-valued Autoregressive Model with Serially Dependent innovations **43**

5 Paper B. Multivariate Mixed Poisson Generalized Inverse Gaussian INAR(1) Regression **64**

6 Paper C. EM Estimation for Bivariate Mixed Poisson INAR(1) Claim Count Regression Models with Correlated Random Effects **85**

7 Paper D. Cluster point processes and Poisson thinning INARMA 115

8 Paper E. INAR Approximation of Bivariate Linear Birth and Death Process 147

9 Concluding Remarks and Future Research 196

9.1 Application of Point Processes 197

9.2 Perspectives on Future Research 198

Bibliography 200

Accompanying Papers

A Zezhun Chen, Angelos Dassios, and George Tzougas
A first-order binomial-mixed Poisson integer-valued autoregressive model with serially dependent innovations
Journal of Applied Statistics (2023), 50:2, 352-369
https://doi.org/10.1080/02664763.2021.1993798

B Zezhun Chen, Angelos Dassios, and George Tzougas
Multivariate mixed Poisson Generalized Inverse Gaussian INAR(1) regression
Comput Stat 38, 955–977 (2023)
https://doi.org/10.1007/s00180-022-01253-0

C Zezhun Chen, Angelos Dassios, and George Tzougas
EM estimation for bivariate mixed poisson INAR(1) claim count regression models with correlated random effects
European Actuarial Journal (2023)
https://doi.org/10.1007/s13385-023-00351-7

D Zezhun Chen, Angelos Dassios
Cluster point processes and Poisson thinning INARMA
Stochastic Processes and their Applications (2022),Volume 147, May 2022, Pages 456-480
https://doi.org/10.1016/j.spa.2022.02.002

E Zezhun Chen, Angelos Dassios, and George Tzougas
INAR approximation of bivariate linear birth and death process
Statistical Inference for Stochastic Process 26, 459–497 (2023)
https://doi.org/10.1007/s11203-023-09289-9

CHAPTER 1

Introduction

1.1 Integer-Valued Time series

Modelling the integer-valued count time-series has attracted a lot of attention over last few years in a plethora of different scientific field such as social sciences, health-care, insurance, economics and financial industry. The standard ARMA model will inevitably introduce real-valued results, and so is not appropriate for modelling count data. As a result, many alternative classes of count time series models have been introduced and explored in the applied statistical literature. An early contribution has been done by Jacobs and Lewis (1978a,b, 1983), who introduced the discrete Autoregressive and Moving average model (DARMA) for stationary discrete time series. However, the correlation structure of DARMA is quite different from the standard time series model. Later, regarding the univariate case Al-Osh and Alzaid (1987) and McKenzie (1985) were the first to consider an INAR(1) model based on the so-called binomial thinning operator. This is introduced as counterpart to the Gaussian AR(1) model for Poisson counts. The idea here is to manipulate the operation between coefficients and variables, as well as innovation term, in such a way that the values are always integers. The relationship of coefficients and variables is defined as $\alpha \circ X = \sum_{i=1}^{X} B_i$ such taht B_i are i.i.d Bernoulli random variables with success probability α and $\circ$ denote the thinning operator. The binomial thinning straightforward to interpret, the probability of survival from the last state. More importantly, compared to DARMA, INAR(1) has the same autocorrelation structure as the standard AR(1) model. One of the popular choices is the INAR(1) with

binomial thinning and Poisson as innovation. The model constructed in this way has Poisson marginal distribution.

In practice, however, the Poisson assumption (variance equals to mean) will be violated for most of the time as the data could be over or under dispersion, could have excessive number of zeros, etc. Consequently, many articles focused on extending this setup by applying different thinning operators or by varying the distribution of innovations to accommodate different features exhibited by count data. Common choices are Bernoulli, Geometric, Poisson, Mixture Poisson, Generalized Inverse Gaussian. For more details, the interested reader can refer to Weiß (2018, 2008b), Davis et al. (2016b),Scotto et al. (2015), among many more.

On the other hand, the literature which focuses on the multivariate case is less developed. In particular, Latour (1997) introduced a multivariate GINAR(p) model with a generalized thinning operator. Karlis and Pedeli (2013) and Pedeli and Karlis (2011, 2013a,b) focused on the diagonal case under which the thinning operators do not introduce cross correlation among different counts. In this case, the dependence structure introduced by innovations. Additionally, Ristić et al. (2012), Popović (2016), Popović et al. (2016) and Nastić et al. (2016)constructed multivariate INAR distributions with cross correlations among counts and random coefficients thinning. Finally, Karlis and Pedeli (2013) extended the setup of the previous articles by allowing for negative cross correlation via a copula-based approach for modelling the innovations.

In count data modelling, mixed Poisson distribution is a common choice as it can accommodate different features arising from data (dispersion, skewness, excess of zeros). However, the autocorrelation for such data is usually ignored as there is no proper model for integer-valued data. Our first three papers (A,B,C) propose models on combining integer-valued time series and mixed Poisson random variables to obtain a more generic framework to model count data.

In Paper A, we develop a new family of binomial-mixed Poisson INAR(1) (BMP INAR(1)) processes by adding a mixed Poisson component to the innovations of the classical Poisson INAR(1) process. Due to the flexibility of the mixed Poisson component, the model includes a large class of INAR(1) processes with different transition probabilities. Moreover, it can capture overdispersion features coming from the data while keeping the innovations serially dependent. We discuss its statistical properties, stationarity conditions and transition probabilities for different mixing densities (Exponential, Lindley). Then, we derive the maximum likelihood estimation method and its asymptotic properties for this model. Finally, we demonstrate our approach using a real data example of iceberg count data from a financial system.

In Paper B, In this paper, we present a novel family of multivariate mixed Poisson-Generalized Inverse Gaussian INAR(1), MMPGIG-INAR(1), regression models for modelling time series of overdispersed count response variables in a versatile manner. The statistical properties associated with the proposed family of models are discussed and we derive the joint distribution of innovations across all the sequences. Finally, for illustrative purposes different members of the MMPGIG-INAR(1) class are fitted to Local Government Property Insurance Fund data from the state of Wisconsin via maximum likelihood estimation.

In Paper C, This article considers bivariate mixed Poisson INAR(1) regression models with correlated random effects for modelling correlations of different signs and magnitude among time series of different types of claim counts. This is the first time that the proposed family of INAR(1) models is used in a statistical or actuarial context. For expository purposes, the bivariate mixed Poisson INAR(1) claim count regression models with correlated Lognormal and Gamma random effects paired via a Gaussian copula are presented as competitive alternatives to the classical bivariate Negative Binomial INAR(1) claim count regression model which only allows for positive dependence between the time series of claim count responses. Our main achievement is that we develop novel alternative Expectation-Maximization type algorithms for maximum likelihood estimation of the parameters of the models which are demonstrated to perform satisfactory when the models are fitted to Local Government Property Insurance Fund data from the state of Wisconsin.

Paper A focus on univariate sequence modelling while Paper B considers a simple, parsimony multivariate integer-valued model for correlated multivariate count data. As the distribution function as well as the log likelihood function becomes more complicated in multivariate INAR model, Expectation-Maximization algorithm is developed for statistical inference for such model which is demonstrated in paper C.

In addition to integer-valued models, there are other classical stochastic process modelling count data as well, e.g. Poisson point process, birth-and-death process. Motivated by Kirchner (2016, 2017), one can carefully construct integer-valued models to approximate Hawkes point processes, and so as other point processes. The rationale and necessity behind this is straightforward: continuous measurement of observations is rare in practice while maximum likelihood estimation method for these stochastic processes rely on continuous observations. Integer-valued time series, by construction, model discrete observations and hence offer an alternative estimation method for these point processes. More importantly, I truly believe that a nice discrete approximation could promote the usage of classical stochastic models (Stochastic volatility model and ARCH model). Before approximating them,

we carefully investigate their mathematical formulation and properties. They are introduced in the following:

1.2 Cluster Point Processes

In insurance modelling, the Poisson process has been used to as a claim arrival process. However, the homogeneous intensity assumption is not realistic in practice. Therefore, an alternative point process was introduced as Cox process, also called doubly stochastic Poisson Processes. For detailed review, see Cox (1955) Bartlett (1963) Serfozo (1972) Bening and Korolev (2012). For application on reinsurance, see Dassios and Jang (2003, 2005, 2008).

The Hawkes Process, which was first introduced by in Hawkes (1971a,b), is a self-exciting point process that its intensity depends on the past of the point process itself. The Hawkes process can be viewed as a contagion (cluster) process in the sense that immigrants arrive as a stationary Poisson process and each immigrant acts as a branching process and generate its offspring (cluster). Due to its simplicity and flexibility, the Hawkes process is applied in different areas, for example seismology in Ogata (1988), epidemiology in Kim (2011), sociology in Mohler et al. (2011), and finance.

However, in some context such as modelling the credit contagion in Jarrow and Yu (2001), the clustering of default is consistent with the Hawkes process, but the default intensity could be impacted exogenously by other factors, which indicates the inappropriateness of homogeneous assumption on immigrant processes. To address this, Dassios and Zhao (2011) introduced the dynamic contagion process by generalizing the Hawkes process with immigrant process as the Cox process, which allows the cluster centers act as a stochastic process.

1.3 Birth and Death Processes

The simple linear birth and death process, which was first introduced by Feller (1939), is a widely used Markov model with applications in population growth, epidemiology, genetics and so on. The basic idea of this process is that the probabilities of any individual giving birth to a new individual, or any individual dying, are constant at any moment in time and all individuals are independent of each other. Many statistical properties, including moments, distribution function, extinction probability, or some other cumulative distribution of interests, are explicitly derived in the

literature; see for example, Kendall (1949). The statistical inference for simple birth and death processes is then developed by Keiding (1975), where maximum likelihood estimators and other asymptotic results are discussed. Since the distribution function of simple birth and death processes is explicit, the construction of the likelihood function is straightforward. However, it is pointed out in the literature that the transition probability is actually cumbersome and numerically unstable when the size of population is large over time. At the same time, a variety of alternative estimation methods have been proposed. For example, quasi- and pseudo - likelihood estimators Chen and Hyrien (2011) Crawford et al. (2014) addressed it as a missing data problem and apply an EM algorithm to maximize it. Tavaré (2018) found those transition probabilities by numerical inversion of the probability generating function and then applied Bayesian methods to perform estimation. Davison et al. (2021) adopted a saddle point approximation method to further improve the accuracy of transition probabilities.

The remaining two papers (D E) carefully construct integer-valued models to approximate classical point processes. We proved that the proposed integer-valued models converge weakly to these point processes. They are summarised in the following:

In Paper D, we consider Poisson thinning Integer-valued time series models, namely integer-valued moving average model (INMA) and Integer-valued Autoregressive Moving Average model (INARMA), and their relationship with cluster point processes, the Cox point process and the dynamic contagion process. We derive the probability generating functionals of INARMA models and compare to that of cluster point processes. The main aim of this paper is to prove that, under a specific parametric setting, INMA and INARMA models are just discrete versions of continuous cluster point processes and hence converge weakly when the length of subintervals goes to zero.

In Paper E, we propose a new type of univariate and bivariate Integer-valued autoregressive model of order one (INAR(1)) to approximate univariate and bivariate linear birth and death process with constant rates. Under a specific parametric setting, the dynamic of transition probabilities and probability generating function of INAR(1) will converge to that of birth and death process as the length of subintervals goes to 0. Due to the simplicity of Markov structure, maximum likelihood estimation is feasible for INAR(1) model, which is not the case for bivariate and multivariate birth and death process. This means that the statistical inference of bivariate birth and death process can be achieved via the maximum likelihood estimation of a bivariate INAR(1) model.

The thesis is organized as follows: in Chapter 2, we show detail definitions of Integer-valued models and our main contributions on this field. In Chapter 3, definitions of point processes are given and we show the way to approximate these point processes by manipulating operators and different discrete random variables in integer-valued models. In the following chapters 4-8 are accompanying papers (A, B, C, D, E, F). The final chapter summarises the overall thesis, discussing potential applications and perspectives for future research.

CHAPTER 2

Integer-value time series

Statistical data arising from many areas are expressed in terms of discrete values, mostly non-negative, e.g. patients in a hospital, employees of a company, item sold from a grocery, insurance claims, trading volume from a finance sector. These data are recorded on a regular basis, for example number of item sold is recorded everyday for a grocery while number of insurance claims are recorded on monthly basis or even yearly basis. Then we have a data set that tracks the sample over time and it forms a time series X_t. If X_t are independent, one can find a discrete random variable to model it and popular choices are Poisson and negative binomial distribution. However, in many cases, these data exhibits autocorrelation and standard time series model (ARIMA) will inevitably introduce real-values and even negative values.

To overcome this, integer-valued time series model was introduced in 1980s and the classical Inter-valued model of order-one (INAR(1)) is defined as:

Definition 2.1. *Let X_t be a non-negative integer-valued time series and ϵ_t be i.i.d discrete random variable with mean μ and finite variance σ^2. The INAR(1) model is defined as*

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \quad \alpha \in [0, 1] \tag{2.1}$$

The $\circ$ is a binomial thinning operator such that $\alpha \circ X = \sum_{i=1}^{X} B_i$ where B_i are i.i.d Bernoulli random variable with success probability α. Then $\alpha \circ X$ as a whole is a binomial random variable.

The INAR(1) model defined in this way simply states that the components of the process X_t are either the survivals from previous time X_t or new 'immigrants' from

the innovations. The binomial thinning operator shares several properties with the multiplication operators,

- Commutativity: changing the order will not change the distribution, i.e.

$$\alpha_1 \circ (\alpha_2 \circ X) \overset{d}{=} \alpha_2 \circ (\alpha_1 \circ X) \overset{d}{=} \alpha_1\alpha_2 \circ X, \quad \alpha_1, \alpha_2 \in [0,1]$$

- Linear with respect to expectation:

$$\mathbb{E}[\alpha_1 \circ X] = \alpha_1 \mathbb{E}[X]$$

- Distributive property,

$$\alpha_1 \circ (X + Y) \overset{d}{=} \alpha_1 \circ X + \alpha_1 \circ Y, \quad X, Y \in \mathbb{N}_0 \text{ and } X \perp\!\!\!\perp Y$$

2.1 Distributional and Statistical Properties

These nice properties of binomial thinning operator enable us to express autoregressive form into moving average form:

$$X_t = \alpha \circ X_{t-1} + \epsilon_t = \alpha^2 \circ X_{t-2} + \epsilon_t + \alpha \circ \epsilon_{t-1} = \alpha^t \circ X_0 + \sum_{j=0}^{t-1} \alpha^j \circ \epsilon_{t-j} \tag{2.2}$$

It implies that the dependence of X_t on the sequence ϵ_t has exponential decay which matches the real-valued AR(1) model. On the other hand, the limiting distribution of X_t is also given by the equation (2.2). Denote the probability generating function of a random variable X as $\Phi_X(\theta) = \mathbb{E}[\theta^X]$. Conditional on the initial variable X_0, the distribution of X_t can be characterized as:

$$\begin{aligned}\Phi_{X_t}(\theta) &= \mathbb{E}[\theta^{\alpha^t \circ X_0}] \prod_{j=0}^{t-1} \mathbb{E}[\theta^{\alpha^j \circ \epsilon_{t-j}}] \\ &= \Phi_{X_0}(1 - \alpha^t + \alpha^t\theta) \prod_{j=0}^{t-1} \Phi_\epsilon(1 - \alpha^j + \alpha^j\theta)\end{aligned}$$

As t goes to infinity, the convergence of this product sequence is guaranteed by two conditions: $\alpha < 1$ and $\mathbb{E}[\epsilon] < \infty$. The convergence indicates existence of limiting distribution of $\Phi_X = \lim_{t\to\infty} \Phi_{X_t}$. Together with the irreducible and aperiodic of X_t,

we can say that the limiting distribution is the unique stationary distribution of X_t. Instead of using infinity product, the recursive relationship in (2.2) leads to another simple expression for probability generating function of stationary distribution of X

$$\Phi_X(\theta) = \Phi_X(1 - \alpha + \alpha\theta)\Phi_\epsilon(\theta) \tag{2.3}$$

This equation is related to the definition of a self-decomposable distribution. For example, if ϵ is a Poisson random variable with rate λ, the probability generating function of X is given by

$$\Phi_X(\theta) = \exp\left\{\frac{\lambda}{1-\alpha}(1-\theta)\right\} \tag{2.4}$$

It is again a Poisson random variable with rate $\frac{\lambda}{1-\alpha}$. Many other well-known discrete distributions belong to the this class, e.g. negative binomial, generalized Poisson, discrete stable distribution.

It is then straightforwardly to derive unconditional moments and stationary moments via above distributional properties.

- Unconditional mean $\mathbb{E}[X_t] = \alpha\mathbb{E}[X_{t-1}] + \mu = \alpha^t\mathbb{E}[X_0] + \mu\sum_{j=0}^{k-1}\alpha^j$
- Unconditional variance

$$\begin{aligned}\mathrm{Var}(X_t) &= \alpha^2\mathrm{Var}(X_{t-1}) + \alpha(1-\alpha)\mathbb{E}[X_{t-1}] + \sigma^2 \\ &= \alpha^{2t}Var(X_0) + (1-\alpha)\sum_{j=1}^{t}\alpha^{2j-1}\mathbb{E}[X_{t-j}] + \sigma^2\sum_{j=1}^{t}\alpha^{2(j-1)}\end{aligned}$$

- Stationary mean $\mathbb{E}[X] = \frac{\mu}{1-\alpha}$. Stationary Variance $\mathrm{Var}(X) = \frac{\alpha\mu+\sigma^2}{1-\alpha^2}$
- Unconditional covariance function $\gamma(k) = \mathrm{Cov}(X_{t-k}, X_t)$

$$\begin{aligned}\gamma(k) &= \mathrm{Cov}(X_{t-k}, \alpha^k \circ X_{t-k}) + \mathrm{Cov}\left(X_{t-k}, \sum_{j=0}^{k-1}\alpha^j \circ \epsilon_{t-j}\right) \\ &= \alpha^k\mathrm{Var}(X_{t-k}) + \sum_{j=0}^{k-1}\alpha^j\mathrm{Cov}(X_{t-k}, \epsilon_{t-j}) \\ &= \alpha^k\gamma(0)\end{aligned}$$

2.2 Parameter estimation and Forecasting

There are a lot of research papers covering the topic on estimating the INAR-type models. For the first order INAR model, the estimation is usually straightforward as X_t is a Markov process and the transition probability is the convolution of innovations and binomial random variable. The likelihood function $\ell(\Theta)$, where Θ is the parameters set of the INAR model, is given by

$$\ell(\Theta) = \prod_{t=1} \left(\sum_{s=0}^{\min\{X_t, X_{t-1}\}} \Pr(\alpha \circ X_{t-1} = X_t - s | X_{t-1}) \Pr(\epsilon_t = s) \right) \tag{2.5}$$

Therefore, maximum likelihood estimation method would be convenient to apply. Apart from that, other classical estimation methods: conditional least square estimation, moments estimation (Yule-Walker equations) as well as their asymptotic properties are also available in literature. It shows that for small value of α e.g. $\alpha < 0.2$, these methods have little difference in terms of bias and MSE while maximum likelihood method yields the best result when $\alpha > 0.5$.

On the other hand, the forecast procedure of INAR model have also been well explored in the literature. The conditional distribution of X_{t+k} given X_t is

$$X_{t+k} = \alpha^k \circ X_t + \sum_{j=0}^{k-1} \alpha^j \circ \epsilon_{t+k-j} \tag{2.6}$$

The classical way to make a prediction is to minimising the L^2 norm which yields a conditional expectation $\mathbb{E}[X_{t+k}|X_t]$ as k-step ahead prediction

$$\hat{X}_{t+k} = \mathbb{E}[X_{t+k}|X_t] = \hat{\alpha}^k X_t + \sum_{j=0}^{k-1} \hat{\alpha}^j \mathbb{E}[\epsilon_{t+k-j}] \tag{2.7}$$

The major concern for this prediction is that it will hardly generate integer-value response which is not a coherent prediction. To maintain the integer-valued nature of this type of model, one can instead minimise the L^1 norm,

$$\hat{X}_{t+h} = \underset{X_{t+h}}{\operatorname{argmin}} \, \mathbb{E}[|X_{t+h} - X_t||X_t] \tag{2.8}$$

i.e. the median of conditional distribution, to obtain a integer-valued prediction. As pointing out by Freeland and McCabe (2004), one need to be careful on this point

estimate, as it may be misleading when the support of the X_{t+h} is small.

2.3 INAR model with correlated innovations (Paper A)

Before constructing the model, we need to introduce a special random variable, mixed Poisson random variable. Most of the time, assuming that the count data are generated from Poisson distribution seems unrealistic as the variance and mean of the data are different from each other significantly. Then a new family of Poisson distribution is purposed by introducing another random variable to model the rate of Poisson distribution.

Formally speaking, there are two definitions and they are different in the way introducing the mixing random variable to the rate parameter.

Definition 2.2. *X is said to follow mixed Poisson distribution if it has following probability mass function*

$$\Pr(X = x) = \int_0^\infty \frac{e^{-\theta}\theta^x}{x!} g(\theta) d\theta \tag{2.9}$$

That is $X \sim Poi(\theta)$ where θ is called mixing random variable with non-negative support and $g(\theta)$ is called mixing density function of θ.

Definition 2.3. *Suppose θ is a non-negative support random variable with unit expectation and density function $g(\theta)$, X is said to follow mixed Poisson distribution if it has following probability mass function*

$$\Pr(X = x) = \int_0^\infty \frac{e^{-\lambda\theta}(\lambda\theta)^x}{x!} g(\theta) d\theta \tag{2.10}$$

That is $X \sim Poi(\lambda\theta)$. The constraint on expectation of the mixing random variable is to avoid identification problem when it comes to statistical inference of ϵ.

The role of mixing random variable looks a bit more clear in the second definition as it controls only the variation of X and λ is always the mean of X. Well-known discrete distribution can be recovered by suitable choice of mixing density, for example geometric, negative binomial random variable are mixed Poisson with exponential

and gamma as mixing density, respectively. Due to its flexibility, mixed Poisson distribution is widely used in count data modelling and one can choose a suitable choice mixing density to adapt the feature of the real data, e.g. under or over dispersion.

In Weiß (2015), the classical Poisson INAR(1) was extended by allowing the innovations ϵ to depend on the current state of the process X_{t-1}, i.e. $\epsilon \sim \text{Po}(aX_{t-1} + b)$ where a and b are some positive constants. The innovation with this definition is separable in the sense that $\varepsilon_t = a * X_{t-1} + \epsilon_t$, where $a * X_{t-1} = \sum_{i=1}^{X_{t-1}} U_i$, with $U_i \overset{i.i.d}{\sim} Po(a)$ and $\epsilon_t \sim Po(b)$. To introduce further heterogeneity while maintaining the model structure, we extend this by allowing U_i to be a mixed Poisson random variable.

Definition 2.4. *The Binomial Mixed Poisson integer-valued Autoregressive model (BMP INAR(1)) is defined by the following equations*

$$\begin{aligned} X_{t+1} &= p_1 \circ X_t + \varepsilon_{t+1} \\ &= p_1 \circ X_t + \varphi *_g X_t + Z_{t+1} \\ p_1 \circ X_t &= \sum_{k=1}^{X_t} V_k, \quad \varphi *_g X_t = \sum_{i=1}^{X_t} U_i \\ P(U_i = x) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^x}{x!} g(\theta_i|\varphi) d\theta_i, \end{aligned} \tag{2.11}$$

- $\circ$ *is a binomial thinning operator such that* V_i *are i.i.d Bernoulli random variables with parameter* $p_1 \in [0, 1]$
- $\{Z_t\}_{t=1,2,\ldots}$ *are i.i.d Poisson random variables with rate* $\lambda_1 > 0$
- $*_g$ *is a reproduction operator such that* U_i *are independent Mixed Poisson distributed with mixing density function* $g(\theta_i|\varphi)$
- $*_g$ *and* $\circ$ *are independent of each other so that* U_i *and* V_k *are independent of each other.*

We extend the classical definition of binomial thinning operator to as a reproduction operator $*_g$. The model can be seen as a population model where binomial part indicates the survivors from the last state, the mixed Poisson part is the total offspring and the innovation part are new immigrants. From statistical inference point of view, it would be nice that each component in equation (2.11) has a explicit distribution function, i.e. we want to make sure the random variable $Y_{t+1} = \varphi *_g X_t$ has explicit

distribution rather than something expressed in integral form. Its distribution is given by

$$\Pr(Y_{t+1} = y|X_t = n) = \mathbb{E}\left[\frac{e^{-\sum_{i=1}^{n}\theta_i}(\sum_{i=1}^{n}\theta_i)^y}{y!}\right], \tag{2.12}$$

where the expectation is taken over all the independent θ_i. We let θ_i be random variable from exponential family so that many of them have the 'additivity' property, i.e. the distribution $\sum_i \theta_i$ is known given the distribution of θ_i. For example, Gaussian, exponential, Gamma, Geometric, Bernoulli and so forth. In paper, we explore two mixing density, exponential and Lindley density.

- In the exponential case, $g(\theta) = \frac{1}{\varphi}e^{-\frac{1}{\varphi}\theta}$, the distribution of U_i and Y_{t+1} are given by

$$\begin{aligned}\Pr(U_i = x) &= \left(\frac{1}{1+\varphi}\right)\left(\frac{\varphi}{1+\varphi}\right)^x, \quad x = 0, 1, \dots \\ \Pr(Y_{t+1} = y|X_t = n) &= \binom{y+n-1}{y}\left(\frac{1}{1+\varphi}\right)^n\left(\frac{\varphi}{1+\varphi}\right)^y\end{aligned} \tag{2.13}$$

- In Lindley case where $g(\theta) = \frac{\varphi^2}{1+\varphi}(\theta+1)e^{-\varphi\theta}$, the distribution of U_i and Y_{t+1} are given by

$$\begin{aligned}\Pr(U_i = x) &= \frac{\varphi^2(\varphi+2+x)}{(1+\varphi)^{x+3}}, \quad x = 0, 1, \dots \\ \Pr(Y_{t+1} = y|X_t) &= \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \sum_{k=0}^{n} C_n^k C_{n+k+y-1}^y (1+\tilde{\varphi})^{-(n+k+y)} \\ \tilde{\varphi} &= \frac{1-\varphi+\sqrt{(\varphi-1)^2+8\varphi}}{2\varphi}\end{aligned} \tag{2.14}$$

In general, the stationary condition for this model given by $p + \mu_g < 1$ where $\mu_g = \mathbb{E}[U_i]$ and the moments are given by following proposition

Proposition 2.1. *Assume $p_1 + \mu_g < 1$. The stationary moments of X_t is given by*

$$\begin{aligned}\mathbb{E}[X_t] = \mu_x &= \frac{\lambda_1}{1-p_1-\mu_g} \\ Var(X_t) = \sigma_x^2 &= \mu_x \frac{1-p_1^2+\sigma_g^2}{1-(p_1+\mu_g)^2} \\ Cov(X_t, X_{t-k}) &= \gamma(k) = (p_1+\mu_g)^k \sigma_x^2.\end{aligned} \tag{2.15}$$

From statistical inference point of view, we adopt the maximum likelihood estimation method as all the components in equation (2.11) has precise distribution func-

tions and likelihood function can be constructed straightforwardly. The asymptotic property of the estimator are given by another proposition

Proposition 2.2. *Suppose we have a random sample* $\{X_1, X_2, \ldots, X_n\}$. *Let* $\mathbf{p} = (p_1, \varphi, \lambda_1)$ *denote the parameters vector for the stationary BMP INAR(1) model. The maximum likelihood estimator* $\hat{\mathbf{p}}$ *has the following asymptotic distribution:*

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \sim N(0, \mathbf{I}^{-1}), \tag{2.16}$$

where

$$\mathbf{H} = \begin{Bmatrix} \ell_{p_1 p_1} & \ell_{p_1 \varphi} & \ell_{p_1 \lambda_1} \\ \ell_{\varphi p_1} & \ell_{\varphi\varphi} & \ell_{\varphi \lambda_1} \\ \ell_{\lambda_1 p_1} & \ell_{p_1 \varphi} & \ell_{\lambda_1 \lambda_1} \end{Bmatrix} \quad \mathbf{I} = -\mathbb{E}[\mathbf{H}] \tag{2.17}$$

Those $\ell_{xy} = \frac{\partial^2 \ell}{\partial x \partial y}$ are second partial derivatives of the log likelihood function of X_t. A simulation study on asymptotic property as well as application on financial count data is conducted in paper and we observed an improvement compared to the model from Weiß (2008a).

2.4 Multivariate INAR model with GIG family (Paper B)

In insurance claim modelling, it is usually the case that one would have a single policy that covers different types of claim. In such case, a multivariate count data modelling is needed to model the joint dynamic. By extending univariate setting, we have the following definition

Definition 2.5. *Let* $\mathbf{X}$ *and* $\mathbf{R}$ *be non-negative integer-valued random vectors in* $\mathbb{R}^m$. *Let* $\mathbf{P}$ *be a diagonal matrix in* $\mathbb{R}^{m \times m}$ *with elements* $p_i \in (0, 1)$. *The multivariate Poisson-Generalized Inverse Gaussian INAR(1) is defined as*

$$\mathbf{X}_t = \mathbf{P} \circ \mathbf{X}_{t-1} + \mathbf{R}_t = \begin{bmatrix} p_1 & 0 & \ldots & 0 & 0 \\ 0 & p_2 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & & p_m \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ \vdots \\ X_{m,t-1} \end{bmatrix} + \begin{bmatrix} R_{1,t} \\ R_{2,t} \\ \vdots \\ R_{m,t} \end{bmatrix} \tag{2.18}$$

- $\circ$ *is binomial thinning operator*
- *Conditional on the last state,* $p_i \circ X_{i,t}$ *and* $p_j \circ X_{j,t}$ *are independent of each other,* $\forall i \neq j$.
- $\{R_{i,t}\}_{i=1,\ldots,m}$ *are mixed Poisson random variables* $Po(\theta_t \lambda_{i,t})$ *with the random effect* θ_t *and distribution function* $G(\theta)$
- *The rate* $\lambda_{i,t}$ *is characterized by its observed covariate* $z_{i,t} \in \mathbb{R}^{a_i \times 1}$ *for some positive integer* a_i *and they are connected through a log link function such that* $\log(\lambda_{i,t}) = z_{i,t}^T \beta_i$

Note that $\{R_{i,t}\}_{i=1,\ldots,m}$ share the same random effect θ_t with mixing distribution $G(\theta)$, which means the dependent structure among $X_{i,t}$ can be controlled by the choice of distribution and its corresponding size of parameters. In this case, the correlation is mainly characterized by mixed Poisson random variable. The choice of θ_t is Generalized inverse Gaussian distribution (GIG) as includes a wide range of distribution like gamma, inverse Gaussian and inverse gamma and it can be useful to capture different level of over dispersion presented in the data. The density of GIG is given by:

$$g(\theta) = \frac{(\psi/\chi)^{\frac{\nu}{2}}}{2K_\nu(\sqrt{\psi\chi})} \theta^{\nu-1} \exp\left\{-\frac{1}{2}\left(\psi\theta + \frac{\chi}{\theta}\right)\right\}, \tag{2.19}$$

where $-\infty \leqslant \nu \leqslant \infty, \psi > 0, \chi > 0$ and $K_\nu(\omega)$ is the modified Bessel function of the third kind of order ν and argument ω such that

$$K_\nu(\omega) = \int_0^\infty z^{\nu-1} \exp\left\{-\frac{1}{2}\omega(z + \frac{1}{z})\right\} dz$$

The simple structure of $\mathbf{X}_t$ ensure the feasibility of maximum likelihood estimation method. To evaluate the transition probability of $\mathbf{X}_t$ given $\mathbf{X}_{t-1}$, it is not straightforward to find out the joint distribution of $\mathbf{R}_{t-1}$ as they are correlated of each other. The joint distribution is given by

$$\begin{aligned} f_\phi(\mathbf{k}, t) &= \mathcal{P}(R_{1,t} = k_1, \ldots, R_{m,t} = k_m) \\ &= \mathbb{E}\left[\mathcal{P}(R_{1,t} = k_1, \ldots, R_{m,t} = k_m | \theta_t)\right] \\ &= \prod_{j=1}^m \frac{\lambda_{j,t}^{k_j}}{k_j!} \int_0^\infty e^{-\theta \sum_{i=1}^m \lambda_{i,t}} \theta^{\sum_{i=1}^m k_i} dG(\theta) \end{aligned} \tag{2.20}$$

In the paper, we investigate following distributions and their moments property which are summarize in the following tables

Table 2.1: Parametrization of mixing density based on GIG density

Mixing density	ψ	χ	ν	Range of parameter
Gamma	2ϕ	0	ϕ	$\phi > 0$
GIG with unit mean	$c\phi$	$\frac{\phi}{c}$	fixed constant in $\mathbb{R}$	$c = \frac{K_\nu(\phi+1)}{K_\nu(\phi)}, \phi > 0$
Inverse Gamma	0	2ϕ	$-\phi - 1$	$\phi > 1$

Table 2.2: Moments for the random effect θ_t. Ex.Kurtosis = Kurtosis - 3

Mixing density $g(\theta)$	Variance	Skenwness	Ex.Kurtosis
Gamma	$\frac{1}{\phi}$	$\frac{2}{\sqrt{\phi}}$	$\frac{6}{\phi}$
Inverse Gaussian	$\frac{1}{\phi^2}$	$\frac{3}{\phi}$	$\frac{15}{\phi^2}$
GIG $\nu = -\frac{3}{2}$	$\frac{1}{\phi}$	$\frac{\phi^{\frac{3}{2}}+\phi^{\frac{1}{2}}}{\phi^2}$	$\frac{3+12\phi+15\phi^2}{\phi^3}$
Inverse Gamma	$\frac{1}{\phi-1}$	$\frac{4\sqrt{\phi-1}}{\phi-2}$	$\frac{6(5\phi-1)}{(\phi-2)(\phi-3)}$

We then further explores the probability functions, maximum likelihood estimation of MINAR-GIG model, prediction as well as application on insurance claim data.

2.5 EM algorithm for Multivariate INAR model (Paper C)

The flexibility of mixed Poisson distribution enable the MINAR(1) to be adaptive to practical data. However, the estimation of such model would not be trivial in some cases. For example, log-normal mixing density. In the paper, we investigate different types of mixing densities and derive their corresponding expectation maximum algorithm. For illustrations, we examine the following bivariate mixing densities.

(a) **Univariate Gamma density**

In this case, the bivariate mixed Poisson regression model shares the same random effect $N_t^{(i)} \sim Pois(\lambda_{i,t}\theta)$ $i = 1, 2$. Denote the mixing density function as $f_\phi(\theta) = f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and it has following expression

$$f_\phi(\theta) = \frac{\phi^\phi}{\Gamma(\phi)}\theta^{\phi-1}e^{-\phi\theta}, \tag{2.21}$$

which has unit mean and variance $\frac{1}{\phi}$. Then the unconditional probability mass function $f_{PG}(\mathbf{k}, t)$ of $\mathbf{N}_t := (N_t^{(1)}, N_t^{(2)})$ can be written down in a closed form

$$\begin{aligned} f_{PG}(\mathbf{k}, t) &= \frac{\lambda_{1,t}}{k_1!}\frac{\lambda_{2,t}}{k_2!}\int_0^\infty e^{-(\lambda_{1,t}+\lambda_{2,t})\theta}\theta^{k_1+k_2} f_\phi(\theta)d\theta \\ &= \frac{\Gamma(\phi+k_1+k_2)}{\Gamma(\phi)\Gamma(k_1+1)\Gamma(k_2+1)}\frac{\phi^\phi \lambda_{1,t}^{k_1}\lambda_{2,t}^{k_2}}{(\phi+\lambda_{1,t}+\lambda_{2,t})^{\phi+k_1+k_2}}. \end{aligned} \tag{2.22}$$

(b) **Bivariate Lognormal density**

Suppose now the random vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$ follows bivariate normal distribution, with mean vector $(-\frac{\phi_1^2}{2}, -\frac{\phi_2^2}{2})$ and covariance matrix Σ

$$\Sigma = \begin{bmatrix} \phi_1^2 & \rho\phi_1\phi_2 \\ \rho\phi_1\phi_2 & \phi_2^2 \end{bmatrix} \tag{2.23}$$

Then the random effect vector $\boldsymbol{\theta} = e^{\boldsymbol{\epsilon}} = (e^{\epsilon_1}, e^{\epsilon_2})$ has Lognormal distribution with unit mean. Denote the density function of $\boldsymbol{\epsilon}$ as f_Σ^N and f_Σ^{LN} for Lognormal density. Then they have the following expressions

$$\begin{aligned} f_\Sigma^N(\boldsymbol{\epsilon}) =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\epsilon_1+0.5\sigma_1^2}{\sigma_1}\right)^2 - 2\rho\left(\frac{\epsilon_1+0.5\sigma_1^2}{\sigma_1}\right)\left(\frac{\epsilon_2+0.5\sigma_2^2}{\sigma_2}\right) + \left(\frac{\epsilon_2+0.5\sigma_2^2}{\sigma_2}\right)^2\right]\right\} \\ f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) =& \frac{1}{\theta_1\theta_2} f_\Sigma^N(\log\boldsymbol{\theta}) = f_\Sigma^{LN}(\boldsymbol{\theta}). \end{aligned}$$

The unconditional distribution $f_{PLN}(\mathbf{k}, t)$ of $\mathbf{N}_t$ is expressed as a double integral

$$\begin{aligned} f_{PLN}(\mathbf{k}, t) &= \int_0^\infty\int_0^\infty \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!} e^{-\lambda_{1,t}\theta_1} e^{-\lambda_{2,t}\theta_2}\theta_1^{k_1}\theta_2^{k_2} f_\Sigma^{LN}(\boldsymbol{\theta})d\theta_1 d\theta_2 \\ &= \int_R\int_R \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!}\exp\{-\lambda_{1,t}e^{\epsilon_1} - \lambda_{2,t}e^{\epsilon_2} + k_1\epsilon_1 + k_2\epsilon_2\} f_\Sigma^N(\boldsymbol{\epsilon})d\epsilon_1 d\epsilon_2. \end{aligned} \tag{2.24}$$

All the double integrals with respect to Lognormal density f_Σ^{LN} can be transformed into double integrals with respect to normal density f_Σ^N so that they can be evaluated by Gauss-Hermite quadrature.

(c) **Gaussian copula paired with Gamma marginals**

Suppose now the random vector $\boldsymbol{\theta}$ is distributed as a meta Gaussian copula such that its marginals are two independent Gamma random variables

with parameter (ϕ_1, ϕ_2) respectively. Define uniform random vector $\mathbf{u} = (F_{\phi_1}(\theta_1), F_{\phi_2}(\theta_2))$. The distribution function $F_{GC}(\boldsymbol{\theta})$ and density function $f_{GC}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} F_{GC}(\boldsymbol{\theta}) &= C_\rho(\mathbf{u}) = F_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \\ f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) &= f_{GC}(\boldsymbol{\theta}) = \frac{f_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))}{f_{sn}(\Phi^{-1}(u_1)) f_{sn}(\Phi^{-1}(u_1))} f_{\phi_1}(\theta_1) f_{\phi_2}(\theta_2) \\ &:= c_\rho(\mathbf{u}) f_{\phi_1}(\theta_1) f_{\phi_2}(\theta_2), \end{aligned} \tag{2.25}$$

where $f_\rho(.,.), F_\rho(.,.)$ are the density function and cumulative distribution of bivariate normal random variable with the following expression

$$f_\rho(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2}\right\}. \tag{2.26}$$

The $\Phi(x)$ is the cdf of standard normal random variable with $\Phi^{-1}(x)$ as its quantile function and $f_{sn}(x)$ is the density function of the standard normal random variable. Finally, $f_{\phi_i}(x)$ and $F_{\phi_i}(x)$ are the pdf and cdf of Gamma density function, for $i = 1, 2$. Then a bivariate Poisson Gamma random vector is constructed as $N_t^{(i)} \sim Pois(\lambda_{i,t}\theta_i), i = 1, 2$ with probability mass function $f_{PGC}(\mathbf{k}, t)$ such that

$$\begin{aligned} f_{PGC}(\mathbf{k}, t) &= \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!} \int_0^\infty \int_0^\infty \exp\{-\lambda_{1,t}\theta_1 - \lambda_{2,t}\theta_2\} \theta_1^{k_1}\theta_2^{k_2} f_{GC}(\theta_1, \theta_2) d\theta_1 d\theta_2 \\ &= \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!} \int_0^1 \int_0^1 e^{-\lambda_{1,t}F_{\phi_1}^{-1}(u_1) - \lambda_{2,t}F_{\phi_2}^{-1}(u_2)} F_{\phi_1}^{-1}(u_1)^{k_1} F_{\phi_2}^{-1}(u_2)^{k_2} c_\rho(u_1, u_2) du_1 du_2 \end{aligned}$$

Then the double integral can be evaluated by Gauss-Legendre quadrature.

In bivariate case of equation (2.18) where $m = 2$, the log likelihood function is defined as:

$$\ell(\Theta) = \sum_{t=1}^{n} \log\left(\sum_{k_1=0}^{s_{1,t}} \sum_{k_2=0}^{s_{2,t}} f_{p_1}(k_1, X_{1,t}) f_{p_2}(k_2, X_{2,t}) f_{\mathbf{R}}(X_{1,t+1} - k_1, X_{2,t+1} - k_2)\right)$$

$$s_{i,t} = \min\{X_{i,t+1}, X_{i,t}\}, \text{for } i = 1, 2$$

Given the observed bivariate sequence $\{\mathbf{X}_t\}_{t=1,\ldots,n}$. Let $Y_{i,t} = p_i \circ X_{i,t-1}$ and $\Theta = \{p_1, p_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi}\}$ be the parameter space for this model. Suppose now we observe the latent variable $\{Y_t\}_{t=1,\ldots n}$ and $\{\boldsymbol{\theta}_t\}_{t=1\ldots n}$, then the complete log likelihood function

becomes

$$\begin{aligned}\ell(\Theta|\mathbf{Y},\boldsymbol{\theta}) \propto & \sum_{t=1}^{n}\sum_{i=1}^{2}(Y_{i,t}\log p_i + (X_{i,t}-Y_{i,t})\log(1-p_i)) \\ & + \sum_{t=1}^{n}\sum_{i=1}^{2}(R_{i,t}\log(\lambda_{i,t}) - \lambda_{i,t}\theta_{i,t}) + \sum_{t=1}^{n}\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta}).\end{aligned} \tag{2.27}$$

Define the following posterior density functions

$$\begin{aligned}\pi_1(\boldsymbol{y}|\Theta^{(j)},\mathbf{X}_t,\mathbf{X}_{t-1}) &= \frac{f_{\mathbf{R}}(\mathbf{X}_{t-1}-\mathbf{y})\prod_{i=1}^{2} f_{p_i}(y_i, X_{i,t-1})}{\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})} \\ \pi_2(\boldsymbol{\theta}|\Theta^{(j)},\mathbf{R}_t) &= \frac{\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t,\mathbf{R}_t)f_{\boldsymbol{\theta}}(\boldsymbol{\phi})}{\int_0^\infty\int_0^\infty \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t,\mathbf{R}_t)f_{\boldsymbol{\phi}}(\boldsymbol{\theta})d\theta_1 d\theta_2}, \\ \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t,\mathbf{k}_t) &= e^{-\lambda_{1,t}\theta_1-\lambda_{2,t}\theta_2}\theta_1^{k_{1,t}}\theta_2^{k_{2,t}}\end{aligned} \tag{2.28}$$

Define the posterior expectations with respect to real-value functions $h(.,.)$

$$\begin{aligned}\mathbb{E}^{(j)}_{\boldsymbol{y},t}[h(\boldsymbol{y})] &= \sum_{y_1=0}^{s_{1,t-1}}\sum_{y_2=0}^{s_{2,t-1}} h(\boldsymbol{y})\pi_1(\boldsymbol{y}|\Theta^{(j)},\mathbf{X}_t,\mathbf{X}_{t-1}) \\ \mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[h(\boldsymbol{\theta})|\mathbf{R}_t] &= \int_0^\infty\int_0^\infty h(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta}|\Theta^{(j)},\mathbf{R}_t)d\theta_1 d\theta_2.\end{aligned} \tag{2.29}$$

- **E-step:** Evaluating the Q function $Q(\Theta;\Theta^{(j)})$ given the the parameters estimated in the j-th iteration,

$$\begin{aligned}Q(\Theta;\Theta^{(j)}) = & \sum_{t=1}^{n}\sum_{i=1}^{2}(y_{i,t}^{(j)}\log p_i + (X_{i,t-1}-y_{i,t}^{(j)})\log(1-p_i)) \\ & + \sum_{t=1}^{n}\sum_{i=1}^{2}(r_{i,t}^{(j)}\log(\lambda_{i,t}) - \lambda_{i,t}\hat{\theta}_{i,t}^{(j)}) + \sum_{t=1}^{n}\mathbb{E}^{(j)}_{\boldsymbol{y},t}[\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta})|\mathbf{R}_t]] \\ y_{i,t}^{(j)} = \mathbb{E}^{(j)}_{\boldsymbol{y},t}[Y_i], & \quad r_{i,t}^{(j)} = X_{i,t}-y_{i,t}^{(j)}, \quad \hat{\theta}_{i,t}^{(j)} = \mathbb{E}^{(j)}_{\boldsymbol{y},t}[\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\theta_i|\mathbf{R}_t]].\end{aligned} \tag{2.30}$$

  After breaking down the log likelihood function, it is obvious that except for the log likelihood contributed by binomial distribution, the rest of the terms are almost the same as that of the Q-function of bivariate mixed Poisson regression model discussed in the last session, which means the updating procedure for $\boldsymbol{\beta}_i, \boldsymbol{\phi}$ will be exactly the same, but we need to evaluate different posterior expectations in this case.

- **M-step:** Similarly, we apply the Newton-Raphson algorithm to update the parameters. Based on the structure of $Q(\Theta;\Theta^{(j)})$, the parameters can be updated separately for binomial part $\boldsymbol{p}$, Poisson part $\boldsymbol{\beta}_i$ and random effect part $\boldsymbol{\phi}$

  - The binomial part can be updated simply as the following gradient function has a unique solution

$$\begin{aligned} g(p_i) &= \frac{\sum_{t=1}^{n} y_{i,t}^{(j)}}{p_i} - \frac{\sum_{t=1}^{n}(X_{i,t-1} - y_{i,t}^{(j)})}{1-p_i} = 0 \\ p_i^{(j+1)} &= \frac{\sum_{t=1}^{n} y_{i,t}^{(j)}}{\sum_{t=1}^{n} X_{i,t-1}}, \quad i = 1,2 \\ y_{i,t}^{(j)} &= \begin{cases} \frac{p_i^{(j)} X_{i,t-1}\mathcal{P}(\mathbf{X}_t - \mathbf{1}_i | \mathbf{X}_{t-1} - \mathbf{1}_i)}{\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}, & X_{i,t} \neq 0 \ and \ X_{i,t-1} \neq 0 \\ 0, & otherwise \end{cases} \\ \mathbf{1}_1 &= (1,0)^T \quad \mathbf{1}_2 = (0,1)^T. \end{aligned} \tag{2.31}$$

  - For the Poisson part, the updating equations are the same with different posterior expectation

$$\begin{aligned} \boldsymbol{\beta}_i^{(j+1)} &= \boldsymbol{\beta}_i^{(j)} - H^{-1}(\boldsymbol{\beta}_i^{(j)}) g(\boldsymbol{\beta}_i^{(j)}), \quad i = 1,2 \\ g(\boldsymbol{\beta}_i^{(j)}) &= \mathbf{Z}_i^T \mathbf{V}_i^{(g)} \quad H(\boldsymbol{\beta}_i^{(j)}) = \mathbf{Z}_i^T \mathbf{D}_i^{(H)} \mathbf{Z}_i \\ \mathbf{V}_i^{(g)} &= \left(\left\{ k_{i,t} - \lambda_{i,t}^{(j)} \hat{\theta}_{i,t}^{(j)} \right\}_{t=1,\ldots,n} \right) \\ \mathbf{D}_i^{(H)} &= \mathrm{diag}\left(\left\{ -\lambda_{i,t}^{(j)} \hat{\theta}_{i,t}^{(j)} \right\}_{t=1,\ldots,n} \right). \end{aligned} \tag{2.32}$$

    Note that when the mixing density $f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ is univariate Gamma, the posterior expectation for $\boldsymbol{\theta}$ has a simple expression

$$\hat{\theta}_t^{(j)} = \hat{\theta}_{1,t}^{(j)} = \hat{\theta}_{2,t}^{(j)} = \frac{\phi^{(j)} + r_{1,t}^{(j)} + r_{2,t}^{(j)}}{\phi^{(j)} + \lambda_{1,t}^{(j)} + \lambda_{2,t}^{(j)}}.$$

  - Similarly, for the random effect part $\boldsymbol{\phi}$,

(a) **Univariate Gamma density**

$$\begin{aligned}
\phi^{(j+1)} &= \phi^{(j)} - \frac{g(\phi^{(j)})}{h(\phi^{(j)})}, \\
g(\phi^{(j)}) &= n(\log \phi^{(j)} - \Psi(\phi^{(j)}) + 1) + \sum_{t=1}^{n} \left(\mathbb{E}_{\boldsymbol{y},t}^{(j)}[\mathbb{E}_{\theta,t}^{(j)}[\log \theta | \mathbf{R}_t]] - \hat{\theta}_t^{(j)} \right) \\
h(\phi^{(j)}) &= n((\phi^{(j)})^{-1} - \Psi^{'}(\phi^{(j)})),
\end{aligned} \tag{2.33}$$

(b) **Bivariate Lognormal**

$$\begin{aligned}
\boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)}) g(\boldsymbol{\phi}^{(j)}) \\
g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)} \left[\frac{\partial \log f_{\Sigma}^{N}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r} | \mathbf{R}_t \right] \right] \\
H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)} \left[\frac{\partial^2 \log f_{\Sigma}^{N}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s} | \mathbf{R}_t \right] \right],
\end{aligned} \tag{2.34}$$

(c) **Gaussian copula paired with Gamma marginals**

$$\begin{aligned}
\boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)}) g(\boldsymbol{\phi}^{(j)}) \\
g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\theta},t}^{(j)} \left[\frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r} | \mathbf{R}_t \right] \right] \\
H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\theta},t}^{(j)} \left[\frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s} | \mathbf{R}_t \right] \right].
\end{aligned} \tag{2.35}$$

Remark This model as well as the EM algorithm can be extent to multivariate case straightforwardly. All the steps and the general form of the formula of the EM algorithm in the multivariate case are exactly the same. The only problem is that it would become cumbersome to evaluate the transition probability $\mathcal{P}(\mathbf{X}_t | \mathbf{X}_{t-1})$ as dimension of $\mathbf{X}_t$ increases. In that case, we need to turn Monte Carlo EM algorithm.

The rest of paper implement the above proposed method on insurance data and compared their predictive performance.

CHAPTER 3

Point Processes

A point process models the occurrences of some events of interest over time. We consider the point process on a real half time $\mathbb{R}_+$ which usually refers to the time t. From counting measure prospective, a point process N on the state space $\mathbb{R}_+$ is a measurable mapping from probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into $(\mathcal{N}^{\#}_{\mathbb{R}_+}, \mathcal{B}(\mathcal{N}^{\#}_{\mathbb{R}_+}))$

$$N : \Omega \mapsto \mathcal{N}^{\#}_{\mathbb{R}_+} \tag{3.1}$$

such that $N(A)$ is an integer-valued random variable for each bounded $A \in \mathcal{B}(R_+)$. $\mathcal{N}^{\#}_{\mathbb{R}_+}$ is the family of all boundedly finite integer-valued measures $\mu \in \mathcal{M}^{\#}_{\mathbb{R}_+}$. For example, let $A = (a, b]$, $a, b \in \mathbb{R}_+$, then the mapping $N(A) : (\Omega, \mathcal{F}) \mapsto \mathbb{N}_0$ defines a random variables that counts the number of events that happen during the time interval $(a, b]$. The distribution of the such point process is actually defined by the joint distribution of finite number of such random variables $N(A_1), \ldots N(A_k)$ for some $k > 0$. Usually, it is convenient to assume that the following

1. The point process is stationary. Mathematically speaking, the joint distribution of

   $$N(A_1 + t), \ldots, N(A_k + t) \tag{3.2}$$

   does not depend on $t > 0$ but only depends on the size of the set $A_1, \ldots, A_k$ (length of the interval). In other words, the event occurrence pattern will not change over time.

2. The point process is simple and crude. This is characterized by the following probabilities:

$$\begin{aligned} \Pr(N(\{t\}) \in \{0,1\}) = 1, \quad \forall t > 0. \\ \Pr(N((0,\Delta]) \geqslant 2) = o(\Delta), \quad \Delta \downarrow 0 \end{aligned} \tag{3.3}$$

That is, there won't be any chance to observe more than one event at any instant time and the probability that more than 1 event occur in interval that is negligibly small.

One important characterization of a point process is its intensity process λ_t. For a stationary point process, the existence of λ_t is guaranteed by the (Khinchin's Existence Theorem, proposition 3.3.I Daley and Vere-Jones (2003)) and it is usually defined as

$$\lambda_t = \lim_{\Delta \downarrow 0} \frac{\mathbb{E}N((t, t+\Delta])}{\Delta} \tag{3.4}$$

It represents the expected number of event occurrence in an unit length interval. In particular, we focus on a family of point processes, called Poisson point processes which is defined by a set of probabilities:

$$\begin{aligned} \Pr(N((t,t+\Delta]) = 1) &= \lambda_t \Delta + o(\Delta) \\ \Pr(N((t,t+\Delta]) = 0) &= 1 - \lambda_t \Delta + o(\Delta) \\ \Pr(N((t,t+\Delta]) \geqslant 2) &= o(\Delta) \end{aligned} \tag{3.5}$$

When $\lambda_t = \mu$ where μ is a fixed constant, the above probabilities yield a well-known homogeneous Poisson process and $N((a,b])$ is a Poisson random variable with rate $\int_a^b \lambda_t dt = (b-a)\mu$. One interesting property of this Poisson process is that if there are n events generated by this process, these n points are uniformly distributed over the interval $(a,b]$. This is usually not the case in practice. Here are some examples:

1. When counting the number of passengers in tube station over weekday, it will be more crowded during the peak hours.

2. The relevant insurance claims are more intense in a certain area when it suffers from severe weathers or natural hazards.

3. The volatility of a financial instrument persists over a certain period.

When the counting data shows a obvious seasonal pattern, or other patterns that can be described by a deterministic function, one can use such function to formulating λ_t, e.g. $\lambda_t = b\sin(t) + c$. In other cases, by generalizing the idea from mixed Poisson distribution, one can introduce a random variable or random process into the intensity process to accommodate different features shown from data. In other words, It is usually straightforward to construct a Poisson point process by formulating the λ_t.

3.1 Cluster Point Process

In the following, we are going to briefly introduce Cox process, Hawkes process, and dynamic contagion process. The term 'cluster' is a way to represent these processes, which can help better understand how these processes work.

Definition 3.1. *The (Marked) Cox process with shot-noise intensity, also called doubly stochastic process, is a cluster point process $N^{(C)}$ with stochastic intensity $\lambda^{(C)}$ such that*

$$\lambda_t^{(C)} = \int_0^t \Upsilon_i f(t - c_i) dN^*(t) = \sum_{i:c_i<t}^{N_t^*} \Upsilon_i f(t - c_i). \tag{3.6}$$

- $N_t^* \equiv \{c_i\}_{i=1,2,...}$ are the arrival times of the Poisson process with the constant rate $\rho > 0$
- $\{\Upsilon_i\}$ are i.i.d externally excited jump sizes, realised at times $\{c_i\}$, with distribution $H(x)$, mean μ_Υ and Laplace transform $\hat{h}(u)$
- $f(u)$ is an Riemann integrable function for any bounded interval in $\mathbb{R}_+$

The general definition of Cox process is given by Definition 6.2.I in Daley and Vere-Jones (2003) where the stochastic part is driven by another random measure. We are particularly interested in this shot-noise version (for the following point processes as well) because we believe that the exogenous random shock Υ_i triggered from N^* are temporary and controlled by $f(.)$. For example, a hail N^* weather happens in a certain area, the severity of this is described by Υ_i and property related insurance claim intensity $\lambda_t^{(C)}$ will increase but decay over time $f()$ as the hail will stop an after a time period.

The cluster representation of this point process is somehow obvious: the cluster centers c_i are generated by N^* which will not be counted into $N^{(C)}$. Then clusters

are then formed by inhomogeneous Poisson processes with intensity $\Upsilon_i \int_0^t f(t-c_i)dt$ for all $i : c_i < t$.

Definition 3.2. *The (Marked) Hawkes process is a self-exciting point process $N^{(H)}$ with stochastic intensity $\lambda^{(H)}$ such that*

$$\lambda_t^{(H)} \;=\; \nu + \int_0^t \chi_i \eta(t-\tau_i) dN^{(H)}(t) \;=\; \nu + \sum_{i:\tau_i<t}^{N_t^{(H)}} \chi_i \eta(t-\tau_i), \tag{3.7}$$

- ν is a positive constant.
- $N_t^{(H)} \equiv \{\tau_i\}_{i=1,2,\dots}$ are the event arrival times of the Hawkes process itself.
- $\{\chi_i\}$ are i.i.d self-exciting jump sizes, realised at times $\{\tau_i\}$, with distribution $G(y)$, mean μ_χ and Laplace transform $\hat{g}(u)$. They are independent of $\{\Upsilon_i\}$
- $f(u)$ is an Riemann integrable function for any bounded interval in $\mathbb{R}_+$

From the integral form of intensity process, this is a clearly 'autoregressive' point process such that the intensity process depends on the trajectory of the point process itself. Apart from that, the likelihood function of this point process is straightforward to construct and there is no latent variable. These features popularize the use of the Hawkes process in practice.

On the other hand, this is a well-known cluster point processes discussed in Hawkes and Oakes (1974). It is easier to understand in terms of a population models. The immigrants (cluster centers) arrive as a homogeneous Poisson process with fixed rate ν. Each immigrant generates a Galton-Waston type branching process with expected branching ratio $\mu_x \int_0^\infty \eta(u)du$. A cluster is then formed by including all the generations (and the immigrant) from the branching process.

Definition 3.3. *The generalized dynamic contagion process is a cluster point process $N^{(DCP)}$, with stochastic intensity $\lambda^{(DCP)}$ such that*

$$\lambda_t^{(DCP)} \;=\; \sum_{i:c_i<t}^{N_t^*} \Upsilon_i f(t-c_i) + \sum_{i:\tau_i<t}^{N_t^{(DCP)}} \chi_i \eta(t-\tau_i), \tag{3.8}$$

where

- $N_t^* \equiv \{c_i\}_{i=1,2,\dots}$ *are the arrival times of the Poisson process with the constant rate* $\rho > 0$

- $N_t^{(DCP)} \equiv \{\tau_i\}_{i=1,2,...}$ *are the arrival times of the generalized dynamic contagion process*
- $\{\Upsilon_i\}$ *are i.i.d externally excited jump sizes, realised at times* $\{c_i\}$*, with distribution* $H(x)$*, mean* μ_Υ *and Laplace transform* $\hat{h}(u)$
- $\{\chi_i\}$ *are i.i.d self-exciting jump sizes, realised at times* $\{\tau_i\}$*, with distribution* $G(y)$*, mean* μ_χ *and Laplace transform* $\hat{g}(u)$*. They are independent of* $\{\Upsilon_i\}$
- $f(u)$ *is an Riemann integrable function for any bounded interval in* $\mathbb{R}_+$
- $\eta(u)$ *is another Riemann integrable function for any bounded interval in* $\mathbb{R}_+$

This is obviously a combination of Hawkes process and Cox process where immigrants no longer arrive uniformly on a fixed interval. Instead, it is driven by a exogenous process and arrive accordingly, for example, financial contagion, credit default events. When a event happens, it will spread like a epidemic models or autoregressive model such that the number of new infections depends proportionally on previous infections.

With the help of piece-wise deterministic Markov theory in Davis (1984), we can find out, at least in a differential equation form, the expectation of $f(t, \lambda_t, N_t)$ for some differentiable function f, e.g. moments, probabilities. For example, the infinitesimal generator of generalized dynamic contagion process $N_t^{(DCP)}$ acting on function $f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right)$ within its domain $\Omega(\mathcal{A})$ is given by

$$\begin{aligned}
&\mathcal{A}f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right) \\
&= \frac{\partial f}{\partial t} + \frac{\partial f}{\partial \lambda} + \rho\left(\int_R f\left(t, \lambda_t^{(DCP)} + x, N_t^{(DCP)}\right) dH(x) - f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right)\right) \\
&+ \lambda_t^{(DCP)}\left(\int_R f\left(t, \lambda_t^{(DCP)} + y, N_t^{(DCP)} + 1\right) dG(y) - f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right)\right)
\end{aligned}$$

where we need to make sure that

$$\begin{aligned}
&\left|\int_R f\left(t, \lambda_t^{(DCP)} + x, N_t^{(DCP)}\right) dH(x) - f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right)\right| < \infty \\
&\left|\int_R f\left(t, \lambda_t^{(DCP)} + y, N_t^{(DCP)} + 1\right) dG(y) - f\left(t, \lambda_t^{(DCP)}, N_t^{(DCP)}\right)\right| < \infty
\end{aligned}$$

To derive moments of $\lambda_t^{(DCP)}$ and $N_t^{(DCP)}$, we can simply set $f = \left(\lambda_t^{(DCP)}\right)^k$ or $f = \left(N_t^{(DCP)}\right)^k$, apply the generator and solve the equations.

3.2 Birth and Death Process

Strictly speaking, birth and death process is not a Poisson point process as the sample path can decrease. However, the way to define such a process is closely related to system of probability equations (3.5). A simple birth and death process is to model the evolution of population Z_t, with a homogeneous birth rate λ, death rate μ and initial population Z_0. At any infinitesimal time Δ, the probabilities that give a new birth, new death, nothing happens and more than one birth or death are

$$\begin{aligned}
&\Pr(Z_{t+\Delta} = n+1 | Z_t = n) = \lambda n\Delta + o(\Delta) \\
&\Pr(Z_{t+\Delta} = n-1 | Z_t = n) = \mu n\Delta + o(\Delta) \\
&\Pr(Z_{t+\Delta} = n | Z_t = n) = 1 - (\lambda+\mu)n\Delta + o(\Delta) \\
&\Pr(|Z_{t+\Delta} - Z_t| \geqslant 2 | Z_t = n) = o(\Delta)
\end{aligned} \tag{3.9}$$

For convenience, denote $\Pr(Z_t = n)$ as $P_n(t)$ and the process is a typical continuous Markov process with above transition probabilities. The transition probabilities is expressed as

$$\begin{aligned}
P_n(t+\Delta) = \lambda(n-1)\Delta P_{n-1}(t) + \mu(n+1)\Delta P_{n+1}(t) \\
+ (1 - (\lambda+\mu)n\Delta)P_n(t) + o(\Delta)
\end{aligned} \tag{3.10}$$

Once we rearrange one of them $P_n(t)$ from the right hand side to the left hand side, divide both sides by Δ and take the limit, we can obtain an ordinary differential equation (ODE) to characterize the simple birth and death process

$$\begin{cases}
\frac{dP_n(t)}{dt} & = \lambda(n-1)P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - (\lambda+\mu)nP_n(t) \\
P_{Z_0}(0) & = 1
\end{cases} \tag{3.11}$$

From equations 3.9 and 3.11, it is not clear that what the distribution looks like for the whole process, although it is straightforward to implement simulation. To explore its distributional property, one can apply a linear transform $\sum_n \theta^n$ on both sides and define $\varphi(t,\theta) = \sum_n \theta^n P_n(t)$, we can get a partial differential equation

(PDE) whose solution φ is the probability generating function of Z_t.

$$\begin{aligned}\frac{\partial \varphi}{\partial t} &= \lambda\theta^2\frac{\partial \varphi}{\partial \theta} + \mu\frac{\partial \varphi}{\partial \theta} - (\lambda+\mu)\theta\frac{\partial \varphi}{\partial \theta} \\ &= (\lambda\theta - \mu)(\theta - 1)\frac{\partial \varphi}{\partial \theta} \\ \varphi(0,\theta) &= \theta^{Z_0}\end{aligned} \tag{3.12}$$

This linear PDE can be solved explicitly

$$\begin{aligned}\varphi(t,\theta) &= \left(1 - \alpha(t) + \alpha(t)\frac{\beta(t)\theta}{1-(1-\beta(t))\theta}\right)^{Z_0} \\ \alpha(t) &= \frac{(\lambda-\mu)e^{(\lambda-\mu)t}}{\lambda e^{(\lambda-\mu)t} - \mu}, \quad \beta(t) = \frac{\lambda-\mu}{\lambda e^{(\lambda-\mu)t} - \mu}\end{aligned} \tag{3.13}$$

1. When $\mu = 0$, this will be a pure birth process, which is formally a point process as the sample path is non-decreasing now. The pure birth process followed negative binomial distribution with size Z_0 and success probability $e^{-\lambda t}$ at any time t

$$P_n(t) = \binom{n-1}{Z_0 - 1} e^{-n\lambda t}(1 - e^{-\lambda t})^n, \quad n \geqslant Z_0 \tag{3.14}$$

2. When $\lambda = 0$, this will be a pure death process which follows binomial distribution with size Z_0 and survival probability $e^{-\mu_t}$ at any time t.

$$P_n(t) = \binom{Z_0}{n} e^{-n\mu t}(1 - e^{-\mu t})^n, \quad 0 \leqslant n \leqslant Z_0 \tag{3.15}$$

3. The simple birth and death process is defined when these two rates are strictly positively and it follows a mixture distribution, i.e. zero-modified geometric random variables.

$$Z_t \sim \sum_{i=1}^{Z_0} B_i(\alpha(t))G_i(\beta(t)), \tag{3.16}$$

where B_i are i.i.d Bernoulli random variables and G_i are i.i.d Geometric random variables with mean $\alpha(t)$ and $\frac{1}{\beta(t)}$, respectively.

3.3 Integer-valued approximation of point processes (Paper D,E)

Motivated by Kirchner (2016), integer-valued model can be a good approximation for those complicated stochastic process and potentially facilitate the statistical inference (non parametric estimation, quasi-likelihood estimation). We show in paper D and E that integer-valued time series model are just discrete version of those cluster point processes and bivariate birth and death process. The approximation made by integer-valued time series in fact improves the interpretability of corresponding stochastic process, e.g. ARCH, GARCH models and stochastic volatility model. More importantly, we usually observe bin-sized count (daily, weekly, monthly counts) in practice rather than exact arrival times for each events,

In the cluster point processes case (Paper D), we extend the work from Kirchner (2016) and construct so-called Integer-valued moving average model Y_n (INMA) and integer-valued autoregressive moving average model Z_n (INARMA), both with infinite orders. The main idea of the approximation is that we utilize the additivity property of independent Poisson random variables: If $X_1 \sim \text{Poi}(\lambda_1)$ and $X_2 \sim \text{Poi}(\lambda_2)$, then $X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$.

Take the Cox process and INMA for example: at any bin-sized interval $A_k = ((k-1)\Delta, k\Delta]$, $k \geqslant 1, \Delta = \frac{T}{n}$, Y_k reports the number of count within the interval and the aggregated process $\sum_k Y_k$ will approximate $N_T^{(C)}$ for a fixed time T.

Definition 3.4. *The stationary Possion thinning INMA(∞) model is defined as*

$$\begin{aligned} Y_n &= \sum_{k=0}^{\infty} \beta_k \circ \xi_{n-k} \\ &= \beta_0 \circ \xi_n + \beta_1 \circ \xi_{n-1} + \cdots + \beta_{n-1} \circ \xi_1, \end{aligned} \tag{3.17}$$

where

- $\beta_k \geqslant 0$ *are some non-negative coefficients*
- ξ_k *are i.i.d and follow* $Pois(\mu)$ *with* $\mu > 0$
- $\{\xi_k\}_{k=\ldots,-2,-1,0} \equiv 0$ *as the process is defined on positive state space* $\mathbb{R}_+$

- *The thinning operator* $\circ$ *is defined as*

$$\beta_k \circ \xi_{n-k} = \sum_{i=1}^{\xi_{n-k}} u_i^{(n,k)}, \quad u_i^{(n,k)} \overset{i.i.d}{\sim} Pois(\beta_k),$$

On the other hand, the aggregated process $S_n = \sum_k^n Y_k$ is a cluster point process such that

$$\begin{aligned} S_n &= \sum_{t=1}^{n} Y_t = \sum_{t=1}^{n}\sum_{k=0}^{t-1} \beta_k \circ \xi_{t-k} = \sum_{t=1}^{n}\sum_{k=0}^{n-t} \beta_k \circ \xi_t \\ &= \sum_{t=1}^{n}\sum_{i=1}^{\xi_t} (u_i^{(t,0)} + u_i^{(t+1,1)} + \cdots + u_i^{(n,n-t)}), \quad u_i^{(t,k)} \sim Pois(\beta_k) \\ &\overset{d}{=} \sum_{t=1}^{n}\sum_{i=1}^{\xi_t} u_i^t, \quad u_i^t \sim Pois(\sum_{k=0}^{n-t} \beta_k). \end{aligned} \tag{3.18}$$

The last equality follows from the independence of the Poisson random variables. It is now clear that the aggregated process S_n is a cluster process such that

- ξ_t generates the cluster centres independently.
- u_i^t is a cluster generated by one of the cluster centre from ξ_t, with the size of cluster (exclude the cluster centre) following $Pois(\sum_{k=0}^{n-t} \beta_k)$

It is clearly that one can approximate Cox process by INMA model via specifying parameter β_i to match intensity process $\lambda_t^{(C)}$.

- $\mu = \rho\Delta t$, $\beta_j = \Upsilon_i f(j\Delta)\Delta, j \geqslant 0$
- Υ_i are i.i.d random variables corresponding to each cluster centre $\xi_{i\Delta}$ arriving at $i\Delta$, with the Laplace transform $\hat{h}(u) = \mathbb{E}[e^{-u\Upsilon_i}]$

Similar approximation procedure applies to $N^{(H)}$ and $N^{(DCP)}$. The main results are shown below.

Definition 3.5. *For* $n > 0$, *let* $\{X_t\}_{t=1,\ldots,n}$,$\{Y_t\}_{t=1,\ldots,n}$ *and* $\{Z_t\}_{t=1,\ldots,n}$ *be the INAR sequence, the INMA sequence and the INARMA sequence with the parametric setting* $\Delta = \frac{T}{n}$, $\alpha_0 = \nu\Delta, \alpha_k = \chi_i\eta(k\Delta)\Delta$ *for* $k > 0$, $\beta_j = \Upsilon_i f(j\Delta)\Delta$ *for* $j \geqslant 0$ *and* $\mu = \rho\Delta$.

Define the following three families of point processes,

$$\begin{aligned} N_n^{(H)}(A) &= \sum_{t:t\Delta\in A} X_t \\ N_n^{(C)}(A) &= \sum_{t:t\Delta\in A} Y_t \\ N_n^{(DCP)}(A) &= \sum_{t:t\Delta\in A} Z_t \end{aligned} \tag{3.19}$$

where A is a bounded set in $\mathcal{B}(\mathbb{R}_+)$ and T is a constant such that $T \geqslant \sup A$. The joint distribution of these point processes are uniquely determined by their p.g.fl.s derived in the section 3.

Theorem 3.1. *Let $N^{(H)}, N^{(C)}, N^{(DCP)}$ be the Hawkes process, the Cox process and the generalized dynamic contagion process. For $n > 0$, let $N_n^{(H)}$, $N_n^{(C)}$ and $N_n^{(DCP)}$ be the point processes defined above. Then we have the following weak convergence results*

$$\begin{aligned} &N_n^{(H)} \xrightarrow{w} N^{(H)} \\ &N_n^{(C)} \xrightarrow{w} N^{(C)} \\ &N_n^{(DCP)} \xrightarrow{w} N^{(DCP)} \quad as\ n \to \infty. \end{aligned} \tag{3.20}$$

On the other case (Paper E), the story is slightly different. The parametrization of integer-valued model is actually quite straightforward in the univariate case as the probability generating function (3.13) already indicates the way to construct such process. The corresponding INAR model is defined as followed

Definition 3.6. *A birth and death INAR(1) model with survival probability $\alpha \in [0,1]$ and birth probability $p \in [0,1]$ is defined as*

$$X_t = p *_1 \alpha \circ X_{t-1}, \tag{3.21}$$

where

- *$\circ$ is the binomial operator*
- *$*_1$ is a geometric (reproduction) operator such that $p *_1 X = \sum_{i=1}^{X} g_i^{(1)}$ with $g_i^{(1)}$ being i.i.d geometric random variable with success probability p whose proba-*

bility mass function is given by

$$P(g_i^{(1)} = k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots,$$

- $p *_1 \alpha \circ X = \sum_{i=1}^{\alpha \circ X} g_i^{(1)}$

The last bullet point shows that $X_t|X_{t-1}$ is a mixture random variable corresponding to 3.16. The interpretation is also clear, i.e. individuals who survive up to time t will give birth to some new individuals. The parametrization of the corresponding INAR model is given by

$$\alpha = \frac{(\lambda - \mu)e^{(\lambda-\mu)\Delta}}{\lambda e^{(\lambda-\mu)\Delta} - \mu}, \quad p = \frac{\lambda - \mu}{\lambda e^{(\lambda-\mu)\Delta} - \mu} \tag{3.22}$$

Moving beyond univariate case, we are interested in approximating the bivariate linear birth and death process. Similar definition to the univariate case, there are two populations who have their own birth and death rates $(\lambda_{11}, \lambda_{22}, \mu_1, \mu_2)$. Furthermore, a newborn of one population can be triggered by the other population. For example the transmission of an infected disease, one population will be healthy people and the other one will be infected people. The increment of a population is given by the other population (one get infected or get recovered). Denote the cross birth rates $\lambda_{12}, \lambda_{21}$. This process is characterized by the following ODE

$$\begin{cases} \frac{dP_{m,n}}{dt} = & (\lambda_{11}(m-1) + \lambda_{21}n)\, P_{m-1,n} + \mu_1(m+1)P_{m+1,n} \\ & + (\lambda_{12}m + \lambda_{22}(n-1))\, P_{m,n-1} + \mu_2(n+1)P_{m,n+1} \\ & - ((\lambda_{11} + \lambda_{12} + \mu_1)m + (\lambda_{21} + \lambda_{22} + \mu_2)n)\, P_{m,n} \\ P_{\mathbf{M}_0}(0) = & 1, \quad M_{1,0}, M_{2,0} \in \mathbb{N}_+ \end{cases} \tag{3.23}$$

After applying linear transform $\sum_m \sum_n \theta^m \phi^n$ on both sides, one can obtain a PDE of the joint probability generating function $\Psi(t, \theta, \phi) = \sum_m \sum_n \theta^m \phi^n P_{mn}(t)$

$$\begin{aligned} \frac{\partial \Psi}{\partial t} &= (\lambda_{11}\theta^2 + \lambda_{12}\theta\phi + \mu_1 - \theta(\lambda_{11} + \lambda_{12} + \mu_1))\frac{\partial \Psi}{\partial \theta} \\ &\quad + (\lambda_{22}\phi^2 + \lambda_{21}\theta\phi + \mu_2 - \phi(\lambda_{21} + \lambda_{22} + \mu_2))\frac{\partial \Psi}{\partial \phi} \end{aligned} \tag{3.24}$$

$$\Psi(0, \theta, \phi) = \theta^{M_{1,0}}\phi^{M_{2,0}}$$

However, the resulting PDE has no closed form solution. A bivariate INAR model, on the other hand, can give a clear representation of the whole process. We define a new bivariate INAR model as follows:

Definition 3.7. *A bivariate birth and death INAR(1) model* $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})^T$ *with survival probability* $\alpha_1, \alpha_2 \in [0,1]$ *and birth probability* $\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22} \in [0,1]$ *is defined as*

$$\begin{aligned} Y_{1,t} &= \beta_{11} *_1 \alpha_1 \circ Y_{1,t-1} + \beta_{21} *_2 Y_{2,t-1} \\ Y_{2,t} &= \beta_{12} *_2 Y_{1,t-1} + \beta_{22} *_1 \alpha_2 \circ Y_{2,t-1}, \end{aligned} \tag{3.25}$$

where

- $\circ$ *is the binomial operator*
- $*_2$ *is another geometric (reproduction) operator different from* $*_1$ *such that* $\beta *_2 X = \sum_{i=1}^{X} g_i^{(2)}$ *with* $g_i^{(2)}$ *being i.i.d geometric random variable whose success probability is* β *. The probability mass function is given by*

$$P(g_i^{(2)} = k) = \beta(1-\beta)^k, \quad k = 0, 1, 2, \ldots,$$

- *Conditional on* $\mathbf{Y}_{t-1}$*, the random variables* $\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}$*,* $\beta_{21} *_2 Y_{2,t-1}$*,* $\beta_{12} *_2 Y_{1,t-1}$ *and* $\beta_{22} *_1 \alpha_2 \circ Y_{1,t-1}$ *are all independent of each other.*

The main contribution (Definition 3 and Theorem 7 in Paper E) is that we managed to find out a parameterization of above bivariate INAR to represent bivariate birth and death process and show that the INAR model converges weakly to the corresponding birth and death process.

CHAPTER 4

Paper A. A First Order Binomial Mixed Poisson Integer-valued Autoregressive Model with Serially Dependent innovations

Abstract

Motivated by the extended Poisson INAR(1), which allows innovations to be serially dependent, we develop a new family of binomial-mixed Poisson INAR(1) (BMP INAR(1)) processes by adding a mixed Poisson component to the innovations of the classical Poisson INAR(1) process. Due to the flexibility of the mixed Poisson component, the model includes a large class of INAR(1) processes with different transition probabilities. Moreover, it can capture some overdispersion features coming from the data while keeping the innovations serially dependent. We discuss its statistical properties, stationarity conditions and transition probabilities for different mixing densities (Exponential, Lindley). Then, we derive the maximum likelihood estimation method and its asymptotic properties for this model. Finally, we demonstrate our approach using a real data example of iceberg count data from a financial system.

4.1 Introduction

Modelling the integer-valued count time series has attracted a lot of attention over the last few years in a plethora of different scientific fields such as the social sciences, healthcare, insurance, economics and the financial industry. The standard ARMA model will inevitably introduce real-valued results, and so is not appropriate for modelling this type of data. As a result, many alternative classes of integer-valued time series models have been introduced and explored in the applied statistical literature. The Integer-valued autoregressive process of order one, abbreviated as INAR(1), was proposed by McKenzie (1985) and Al-Osh and Alzaid (1987) as a counterpart to the Gaussian AR(1) model for Poisson counts. This model was derived by manipulating the operation between coefficients and variables, as well as the innovation term, in such a way that the values are always integers. The relationship of coefficients and variables is defined as $\alpha \circ X_t = \sum_{i=1}^{k} V_i$ such that V_i are i.i.d Bernoulli random variables with parameter α and $\circ$ denotes the binomial thinning operator. The binomial thinning is very easy to interpret, and binomial INAR(1) has the same autocorrelation structure as the standard AR(1) model and hence can be applied to fit the count data. For a general review, please see Weiß (2008b) and Scotto et al. (2015).

Later on, in order to accommodate different features exhibited by count data, for example, under-dispersion, overdispersion, probability of observing zero and different dependent structures, many research studies introduced alternative thinning operators or varied the distribution of V_i for different needs. The case where V_i are i.i.d geometric random variables is analyzed by Ristić et al. (2009), which is called NGINAR(1). Kirchner (2016) introduced reproduction operators so that V_i are i.i.d Poisson random variables to explore the relationship between Hawkes process and integer-valued time series. For further variation, random coefficients thinning is introduced so that Vi are i.i.d Bernoulli with the parameter α being a random variable. This type of thinning operator was proposed by McKenzie (1985, 1986) and Zheng et al. (2007); they applied this to a generalized INAR(1) model. In particular, to accommodate the overdispersion feature, one way is to change the thinning operators from binomial to other types as discussed above.Another way is to replace the innovation distribution by some other overdispersed distribution; for example, see Bourguignon et al. (2019). A third approach would be to keep the structure of binomial INAR(1) but to allow the innovation terms to be serially dependent; see Weiß (2015).

In this study, motivated by Weiß (2015), we develop a new family of binomial-mixed Poisson INAR(1) (BMP INAR(1)) processes by adding a mixed Poisson component

to the innovations term of the classical Poisson INAR(1) process. The proposed class of BMP INAR(1) processes is ideally suited for modelling heterogeneity in count time series data since, due to the mixed Poisson component which we introduce herein, it includes many members with different transition probabilities that can adequately capture different levels of overdispersion in the data while keeping the innovation as independent Poisson.

The rest paper is organized as follows. Section 4.2 defines the Binomial mixed Poisson INAR(1) model by adding a mixed Poisson component in the Poisson INAR(1) model. Statistical properties and the stationarity condition are derived in Section 4.3. Section 4.4 derives the distribution of the mixed Poisson component based on two different mixing density functions from the exponential family, namely the Exponential and Lindley distributions. In Section 4.5, maximum likelihood estimation is discussed as well as its asymptotic properties for the estimators. In Section 4.6, the model is fitted to financial data (iceberg count) and discuss numerical results. Finally, concluding remarks are provided in Section 4.7.

4.2 Construction of Binomial Mixed Poisson INAR(1)

In Weiß (2015), the classical Poisson INAR(1) was extended by allowing the innovations ε to depend on the current state of the model X_t such that $\varepsilon_t \sim Po(aX_{t-1}+b)$ where a and b are some positive constants. The innovation with this definition is separable in the sense that $\varepsilon_t = a * X_{t-1} + \epsilon_t$, where $a * X_{t-1} = \sum_{i=1}^{X_{t-1}} U_i$, with $U_i \overset{i.i.d}{\sim} Po(a)$ and $\epsilon_t \sim Po(b)$. To introduce further heterogeneity while maintaining serially dependent innovations structure in this model, we extend this by allowing U_i to be a mixed Poisson random variable.

Starting from a Poisson random variable U with parameter θ, we may obtain a large class of random variables by allowing θ to be another random variable which follows some classes of density function $g(\theta|\varphi)$ where φ can be a scalar or a vector; see Karlis (2005). The random variable U follows a Mixed Poisson distribution with g as a mixing density. The distribution function of U is defined as

$$P(U=u) = \int_0^\infty \frac{e^{-\theta_i}\theta_i^u}{u!} g(\theta|\varphi)d\theta. \tag{4.1}$$

We now construct our model.

Definition 4.1. *The Binomial Mixed Poisson integer-valued Autoregressive model (BMP INAR(1)) is defined by the following equations*

$$\begin{aligned} X_{t+1} &= p_1 \circ X_t + \varepsilon_{t+1} \\ &= p_1 \circ X_t + \varphi *_g X_t + Z_{t+1} \\ p_1 \circ X_t &= \sum_{k=1}^{X_t} V_k, \quad \varphi *_g X_t = \sum_{i=1}^{X_t} U_i \\ P(U_i = x) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^x}{x!} g(\theta_i|\varphi) d\theta_i, \end{aligned} \tag{4.2}$$

where

- $\circ$ *is a binomial thinning operator such that* V_i *are i.i.d Bernoulli random variables with parameter* $p_1 \in [0, 1]$
- $\{Z_t\}_{t=1,2,\ldots}$ *are i.i.d Poisson random variables with rate* $\lambda_1 > 0$
- $*_g$ *is a reproduction operator such that* U_i *are independent Mixed Poisson distributed with mixing density function* $g(\theta_i|\varphi)$
- $*_g$ *and* $\circ$ *are independent of each other so that* U_i *and* V_k *are independent of each other.*

As we will see shortly, the stationarity condition for this model is simply $p_1 + \mu_g < 1$ where μ_g is the first moment of U_i. When it comes to interpretation, this model can be seen as the evolution of a population where the binomial part indicates the survivors from the previous period, the mixed Poisson part is the total offspring and the innovation part indicates immigrants. Obviously, this model is a Markov Chain and its transition probability can be found easily once we know the mixing density $g(\theta|\varphi)$. The probability mass function of $Y_{t+1} = \varphi *_g X_t$ is given by

$$P(Y_{t+1} = y|X_t = n) = \mathbb{E}\left[\frac{e^{-\sum_{i=1}^n \theta_i}(\sum_{i=1}^n \theta_i)^y}{y!}\right], \tag{4.3}$$

where the expectation is taken over $\theta_1, \theta_2, \ldots, \theta_n$. In order to evaluate the expectation explicitly, it would be desirable that the random variables θ_i have an 'additivity' property such that density (or probability mass function) of the sum $\sum_{i=1}^n \theta_i$ is either itself with different parameters or can be written in a closed form. Many members of the exponential family have this kind of property. In general, we let $g(x|\varphi)$ be of

an exponential family form such that

$$g(x|\varphi) = h(x)\exp\{\eta(\varphi)T(x) + \xi(\varphi)\}. \tag{4.4}$$

Denote the density of the sum $S_n = \sum_{i=1}^{n}\theta_i$ as $g_n(s|\varphi)$, where θ_i are i.i.d random variables with density $g(\theta|\varphi)$. The expectation above can be expressed as

$$P(Y_{t+1} = y|X_t = n) = \int_{\mathbb{R}^+} \frac{e^{-s}s^y}{y!} g_n(s|\varphi)ds. \tag{4.5}$$

The density $g_n(s|\varphi)$ is explicitly known in many cases, for example, it can be an Inverse Gaussian, Exponential, Gamma, Geometric, Bernoulli or Lindley. For the sake of parsimony, we use distributions with a single parameter. In other words, we assume that φ is scalar. Note that, if we let $g(\theta|\varphi) = \delta_\varphi(\theta)$ - a Dirac delta function concentrating at φ, the model will recover to the Extended Poisson INAR(1) in Weiß (2015).

4.3 Statistical properties of BMP INAR(1)

4.3.1 Moments and correlation structure

We first need to derive the moments of U_i

Lemma 4.1. *The first moment and second central moment of U_i with density $g(x|\varphi)$ are given by*

$$\mathbb{E}[U_i] = \mu_g, \quad Var(U_i) = \mu_g + \sigma_g^2, \tag{4.6}$$

where $\mu_g = \mathbb{E}_g[\theta_i] = \int_R xg(x|\varphi)dx$ and $\sigma_g^2 = Var_g(\theta_i)$.

Proof. By the conditional expectation argument

$$\mathbb{E}[U_i] = \mathbb{E}_g[\mathbb{E}[U_i|\theta_i]] = \mathbb{E}_g[\theta_i] = \mu_g$$

$$\mathbb{E}[U_i^2] = \mathbb{E}_g[\mathbb{E}[U_i^2|\theta_i]] = \mathbb{E}_g[\theta_i^2 + \theta_i]$$

$$Var(U_i) = \mathbb{E}[U_i^2] - (\mathbb{E}[U_i])^2 = \sigma_g^2 + \mu_g.$$

□

Proposition 4.1. *Assume $p_1 + \mu_g < 1$. The stationary moments of X_t is given by*

$$\begin{aligned} \mathbb{E}[X_t] = \mu_x &= \frac{\lambda_1}{1 - p_1 - \mu_g} \\ Var(X_t) = \sigma_x^2 &= \mu_x \frac{1 - p_1^2 + \sigma_g^2}{1 - (p_1 + \mu_g)^2} \\ Cov(X_t, X_{t-k}) &= \gamma(k) = (p_1 + \mu_g)^k \sigma_x^2. \end{aligned} \tag{4.7}$$

Proof. For the first moment, we have

$$\begin{aligned} \mathbb{E}[X_t] &= \mathbb{E}[p_1 \circ X_{t-1}] + \mathbb{E}[\varphi *_g X_{t-1}] + \mathbb{E}[Z_t] \\ \mu_x &= p_1 \mu_x + \mu_g \mu_x + \lambda_1 \\ \mu_x &= \frac{\lambda_1}{1 - p_1 - \mu_g}. \end{aligned}$$

Since the operators $\circ$ and $*_g$ are independent of each other, for the second central moment, we have

$$\begin{aligned} Var(X_t) &= Var(p_1 \circ X_{t-1} + \varphi *_g X_{t-1}) + Var(Z_t) \\ &= Var(\mathbb{E}[\sum_{i=1}^{X_{t-1}} (V_i + U_i) | X_{t-1}]) + \mathbb{E}[Var(\sum_{i=1}^{X_{t-1}} (V_i + U_i) | X_{t-1})] + \lambda_1 \\ &= (p_1 + \mu_g)^2 \sigma_x^2 + (p_1(1 - p_1) + \sigma_g^2 + \mu_g)\mu_x + \lambda_1 \\ \sigma_x^2 &= \mu_x \frac{1 - p_1^2 + \sigma_g^2}{1 - (p_1 + \mu_g)^2}. \end{aligned}$$

Let $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \dots,)$ be the σ-algebra generated by the model X_t up to time t, the covariance of the model is given by

$$Cov(X_t, X_{t-k}) = Cov(p_1 \circ X_{t-1}, X_{t-k}) + Cov(\varphi *_g X_{t-1}, X_{t-k}) + Cov(Z_t, X_{t-k}).$$

Again by using conditional expectations, we have

$$\begin{aligned} Cov(p_1 \circ X_{t-1}, X_{t-k}) &= Cov(\mathbb{E}[p_1 \circ X_{t-1} | \mathcal{F}_{t-1}], \mathbb{E}[X_{t-k} | \mathcal{F}_{t-1}]) + E[Cov(p_1 \circ X_{t-1}, X_{t-k} | \mathcal{F}_{t-1})] \\ &= Cov(p_1 X_{t-1}, X_{t-k}) + \mathbb{E}[Cov(\sum_{i=1}^{X_{t-1}} V_i, X_{t-k} | \mathcal{F}_{t-1})] \\ &= p_1 \gamma(k-1) + 0. \end{aligned}$$

Obviously, $Cov(X_t, X_{t-k}) = \gamma(k) = (p_1 + \mu_g)\gamma(k-1) = (p_1 + \mu_g)^k\gamma(0)$. □

From the results above, it is clear that this model follows the same correlation structure as that of standard AR(1) model. Furthermore, unlike equal-dispersed Poisson INAR(1), BMP INAR(1) is in general an overdispersed model with Fisher index of dispersion

$$\mathrm{FI}_x = \frac{\sigma_x^2}{\mu_x} = 1 + \frac{\mu_g^2 + 2p_1\mu_g + \sigma_g^2}{1 - (p_1 + \mu_g)^2}. \tag{4.8}$$

4.3.2 Existence of Stationary Solution

Proposition 4.2. *Given that* $P(U_i = 0) > 0$ *and* $p_1 + \mu_g < 1$ *the following infinite sequence*

$$\begin{aligned} f_i(\theta) &= (1 - p_1 + p_1 f_{i-1}(\theta))\Phi_u(f_{i-1}(\theta)), \quad i \geqslant 1 \\ f_0(\theta) &= \theta, \quad \theta \in [0, 1] \end{aligned} \tag{4.9}$$

where $\Phi_u(\theta)$ *is the probability generating function (p.g.f) of* U_i*, and* $\lim_{i\to\infty} f_i(\theta) = 1$

Proof. Define the increment of the sequence

$$\begin{aligned} f_i(\theta) - f_{i-1}(\theta) &= (1 - p_1 + p_1 f_{i-1}(\theta))\Phi_u(f_{i-1}(\theta)) - f_{i-1}(\theta) \\ &= (1 - p_1 + p_1 x)\Phi_u(x) - x \quad x = f_i(\theta) \\ &=: Q(x) \end{aligned}$$

By the definition of p.g.f, $x \in [0, 1]$, the monotonicity of this function is shown by its first and second derivatives

$$\begin{aligned} Q'(x) &= p_1\Phi_u(x) + (1 - p_1 + p_1 x)\Phi_u(x) - 1 \\ Q''(x) &= 2p_1\Phi_u'(x) + (1 - p_1 + p_1 x)\Phi_u''(x) \end{aligned}$$

By the definition of p.g.f, $\Phi'(x) \geqslant 0$ and $\Phi''(x) \geqslant 0$. So $Q''(x) \geqslant 0$, which implies $Q'(x)$ is non-decreasing function. Then we have

$$Q'(x) \leqslant Q'(1) = p_1 + \mu_g - 1 < 0$$

Notice that $Q(0) = (1-p_1)P(U_i = 0) > 0,\ Q(1) = 0$. Hence we can conclude that Q is a monotonic decreasing function ranging from 0 to $Q(0)$. In order words, for any $i = 1, \ldots,$ and $\theta \in [0,1]$, the sequence $f_i(\theta) = f_{i-1}(\theta) + Q(f_{i-1}(\theta))$ is increasing with respect to i. Finally, $\lim_{i\to\infty} f_i(\theta) = 1$ □

Proposition 4.3. *Let X_t be the BMP INAR(1) model defined in 4.1. If the condition $P(U_i) > 0$ and $p_1 + \mu_g < 1$ holds, then the process X_t has a proper stationary distribution and X_t is an ergodic Markov Chain. The stationary distribution is $\Phi_x(\theta) = \prod_{i=0}^{\infty} \Phi_z(f_i(\theta))$*

Proof. Denote the p.g.f of X_n and the innovation Z_n as $\Phi_{X_n}(\theta)$ and $\Phi_z(\theta)$respectively, then $\Phi_{X_n}(\theta)$ can be expressed as following product

$$\begin{aligned}
\Phi_{X_n}(\theta) &= \mathbb{E}[\mathbb{E}[\theta^{X_n}|X_{n-1}]|X_0] \\
&= \mathbb{E}[\mathbb{E}[\theta^{p_1 \circ X_{n-1} + \varphi *_g X_{n-1} + Z_t}|X_{n-1}]|X_0] \\
&= \mathbb{E}[f_1(\theta)^{X_{n-1}}|X_0]\Phi_z(f_0(\theta)) \\
&= \vdots \qquad \vdots \\
&= \mathbb{E}[f_n(\theta)^{X_0}] \prod_{i=0}^{n-1} \Phi_z(f_i(\theta)),
\end{aligned}$$

To show the existence of the limiting distribution is equivalent to show the limit of the product as n goes to infinity is something other than 0, which means that we have to show that the series

$$LP_n = \log \Phi_{X_n}(\theta) = \log \mathbb{E}[f_n(\theta)^{X_0}] + \sum_{i=0}^{n-1} \log \Phi_z(f_i(\theta)),$$

is convergent as $n \to \infty$. The convergence of the infinite series $\sum_{i=0}^{\infty} \log \Phi_z(f_i(\theta))$ can be shown by the ratio test

$$\begin{aligned}
&\lim_{i\to\infty} \left| \frac{\log \Phi_z(f_i(\theta))}{\log \Phi_z(f_{i-1}(\theta))} \right| \\
=&\lim_{x\to 1} \frac{\log \Phi_z((1-p_1+p_1x)\Phi_u(x))}{\log \Phi_z(x)} \\
=&\lim_{x\to 1} \frac{\Phi_z(x)}{\Phi_z'(x)} \frac{\Phi_z'((1-p_1+p_1x)\Phi_u(x))}{\Phi_z((1-p_1+p_1x)\Phi_u(x))} (p_1\Phi_u(x) + (1-p_1+p_1x)\Phi_u'(x)) \\
=&p_1 + \mu_g < 1
\end{aligned} \tag{4.10}$$

Hence $\lim_{n\to\infty} LP_n > -\infty$, from which we can infer that $\lim_{n\to\infty} \Phi_{X_n}(\theta) > 0$ exists and the limiting distribution of X_n exists. Furthermore, by the construction of X_n, the chain is defined on a countable state space $\mathcal{S} = \{0, 1, 2, \dots\}$. The positivity of transition probability $\mathcal{P}(X_n = j | X_{n-1} = i) > 0, \ \forall \ i, j \in \mathcal{S}$ implies that X_n is irreducible and aperiodic. Hence the limiting distribution $\Phi_x(\theta) = \lim_{n\to\infty} \Phi_{X_n}(\theta)$ is the unique stationary distribution for X_n. □

In general, $P(U_i = 0) = \int_{\mathbb{R}^+} e^{-\theta} g(\theta|\varphi) d\theta > 0$ as long as $g(\theta|\varphi) > 0$, so we just need to ensure the existence of the first moment to achieve the stationarity of X_n. The infinite product $\Phi_x(\theta) = \prod_{i=0}^{\infty} \Phi_z(f_i(\theta))$ is the p.g.f of the stationary distribution, which also satisfies

$$\Phi_x(\theta) = \Phi_x\left((1 - p_1 + p_1\theta)\Phi_u(\theta)\right) \Phi_z(\theta). \tag{4.11}$$

4.4 Distribution function of the Mixed Poisson Component

In order to apply maximum likelihood estimation for the statistical inference of this model, we need to derive the distribution of $Y_{t+1} = \varphi *_g X_t$ according to different density functions g. As mentioned before, we focus on the density g coming from the exponential family. For expository purposes, we will derive the distribution of Y_{t+1} based on exponential and Lindley densities.

4.4.1 Mixed by Exponential density

If $g(\theta|\varphi) = \frac{1}{\varphi} e^{-\frac{1}{\varphi}\theta}$, then the distribution of U_i is given by

$$\begin{aligned} P(U_i = x) &= \int_0^{\infty} \frac{e^{-\theta_i}\theta_i^x}{x!} \frac{1}{\varphi} e^{-\frac{1}{\theta_i}x} d\theta_i \\ &= \frac{1}{\varphi x!} \int_0^{\infty} e^{-(1+\frac{1}{\varphi})\theta_i} \theta_i^x d\theta_i \\ &= \left(\frac{1}{1+\varphi}\right)\left(\frac{\varphi}{1+\varphi}\right)^x, \quad x = 0, 1, \dots \end{aligned} \tag{4.12}$$

which is a geometric distribution with parameter $\frac{\varphi}{1+\varphi}$. Then, the distribution function $f_\varphi(m, X_t)$ of $\varphi *_g X_t$ as well as its first and second derivatives are given by

$$\begin{aligned}
f_\varphi(m, X_t) &= C^m_{m+X_t-1}(\frac{1}{1+\varphi})^{X_t}(\frac{\varphi}{1+\varphi})^m \\
\frac{\partial f_\varphi(m, X_t)}{\partial \varphi} &= \left(\frac{m}{\varphi(1-\varphi)} - \frac{X_t}{1+\varphi}\right) f_\varphi(m, X_t) \\
\frac{\partial^2 f_\varphi(m, X_t)}{\partial(\varphi)^2} &= \left(\left(\frac{m}{\varphi(1-\varphi)} - \frac{X_t}{1+\varphi}\right)^2 + \frac{X_t}{(1+\varphi)^2} - \frac{m(1+2\varphi)}{\varphi^2(1+\varphi)^2}\right) f_\varphi(m, X_t).
\end{aligned} \tag{4.13}$$

Note that X_t will recover to the NGINAR(1) in Ristić et al. (2009) if we further let $p_1 = 0$. In general, the stationarity condition becomes $p_1 + \varphi < 1$ and the probability generating function of X_t satisfies the equation

$$\Phi_x(\theta) = \Phi_x\left(\frac{1 - p_1 + p_1\theta}{1 + \varphi - \varphi\theta}\right)\Phi_z(\theta). \tag{4.14}$$

We will now relax the assumption of the innovation term being Poisson and let the marginal distribution of X be a geometric random variable with parameter $\frac{\alpha}{1+\alpha}, \alpha > 0$. Using the relationship of the p.g.f, we can infer the required distribution of Z.

Proposition 4.4. *If $p_1 > \varphi, \alpha > \varphi$ or $p_1 < \varphi, \alpha < \varphi$ and the distribution of $\{Z_t\}_{t=1,2,\ldots}$ follows a mixed geometric distribution such that*

$$Z_t = \begin{cases} Geom(\frac{\varphi}{1+\varphi}), & W.P. \quad \frac{(p_1-\varphi)\alpha}{\alpha-\varphi} \\ Geom(\frac{\alpha}{1+\alpha}), & W.P. \quad 1 - \frac{(p_1-\varphi)\alpha}{\alpha-\varphi} \end{cases}, \tag{4.15}$$

then the marginal distribution of X follows a $Geom(\frac{\alpha}{1+\alpha})$ distribution.

Proof. By utilizing equation 4.14, we assume the X has a geometric distribution such that $\Phi_x(\theta) = \frac{1}{1+\alpha-\alpha\theta}$. Then, the probability generating function of Z has the following form,

$$\begin{aligned}
\Phi_z(\theta) &= \frac{\Phi_x(\theta)}{\Phi_x\left(\frac{1-p_1+p_1\theta}{1+\varphi-\varphi\theta}\right)} \\
&= \frac{(1+\varphi-\varphi\theta)(1+\alpha) - \alpha(1-p_1+p_1\theta)}{(1+\alpha-\alpha\theta)(1+\varphi-\varphi\theta)} \\
&= \frac{(p_1-\varphi)\alpha}{\alpha-\varphi}\frac{1}{1+\varphi-\varphi\theta} + \left(1 - \frac{(p_1-\varphi)\alpha}{\alpha-\varphi}\right)\frac{1}{1+\alpha-\alpha\theta}.
\end{aligned} \tag{4.16}$$

□

4.4.2 Mixed by Lindley density

Suppose now the density $g(\theta|\varphi) = \frac{\varphi^2}{1+\varphi}(\theta+1)e^{-\varphi\theta}$ is a Lindley density function. The distribution of U_i is the so-called Poisson-Lindley distribution, see Karlis (2005), which has the following probability mass function

$$\begin{aligned} P(U_i = x) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^x}{x!}\frac{\varphi^2}{1+\varphi}(\theta_i+1)e^{-\varphi\theta_i}d\theta_i \\ &= \frac{\varphi^2}{(1+\varphi)x!}\left(\int_0^\infty \theta_i^{x+1}e^{-(\varphi+1)\theta_i}d\theta_i + \int_0^\infty \theta_i^x e^{-(\varphi+1)\theta_i}d\theta_i\right) \\ &= \frac{\varphi^2}{(1+\varphi)x!}\left(\frac{\Gamma(x+2)}{(1+\varphi)^{x+2}} + \frac{\Gamma(x+1)}{(1+\varphi)^{x+1}}\right) \\ &= \frac{\varphi^2(\varphi+2+x)}{(1+\varphi)^{x+3}}, \quad x = 0, 1, \dots \end{aligned} \tag{4.17}$$

Under this parameter setting, $\mathbb{E}[U_i] = \mu_g = \frac{\varphi+2}{\varphi(\varphi+1)}$ which makes the parameter φ less interpretable. So we adopt the following parameter setting for the mixing density $g(\theta|\varphi)$

$$g(\theta|\varphi) = \frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}(\theta+1)e^{-\tilde{\varphi}\theta} \quad \tilde{\varphi} = \frac{1-\varphi+\Delta}{2\varphi} \quad \Delta = \sqrt{(\varphi-1)^2+8\varphi} \tag{4.18}$$

Then, $\mu_g = \varphi$, $\sigma_g = \varphi^2 - \frac{2}{(\tilde{\varphi}(1+\tilde{\varphi}))^2}$. On the other hand, the additivity of U_i is not that clear. In order to evaluate the expectation 4.3, we need to find out the distribution of $S_n = \sum_{i=1}^n \theta_i$.

Proposition 4.5. *Suppose θ_i are i.i.d Lindley distributed. The density of the sum $S_n = \sum_{i=1}^n \theta_i$ is given by*

$$g_n(s|\varphi) = \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n e^{-\tilde{\varphi}s}\sum_{k=0}^n \frac{C_n^k}{\Gamma(n+k)}s^{n+k-1}. \tag{4.19}$$

Proof. We can prove this by inverting the Laplace transform. The Laplace transform of θ_i is

$$\begin{aligned} \mathbb{E}[e^{-\nu\theta_i}] &= \int_0^\infty \frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}(\theta_i+1)e^{-(\nu+\tilde{\varphi})\theta_i}d\theta_i \\ &= \frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\frac{\tilde{\varphi}+\nu+1}{(\tilde{\varphi}+\nu)^2}. \end{aligned}$$

Then the Laplace transform of S_n is simply the product of $\mathbb{E}[e^{-\nu\theta_i}]$, which is

$$\mathbb{E}[e^{-\nu S_n}] = \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \frac{(\tilde{\varphi}+\nu+1)^n}{(\tilde{\varphi}+\nu)^{2n}}.$$

Using a binomial expansion, we have

$$\begin{aligned}
\mathbb{E}[e^{-\nu S_n}] &= \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \frac{1}{(\tilde{\varphi}+\nu)^{2n}} \sum_{k=0}^{n} C_n^k (\tilde{\varphi}+\nu)^k \\
&= \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \frac{1}{(\tilde{\varphi}+\nu)^{n}} \sum_{k=0}^{n} C_n^{n-k} (\tilde{\varphi}+\nu)^{-(n-k)} \\
&= \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \sum_{k=0}^{n} C_n^k (\tilde{\varphi}+\nu)^{-(n+k)} \\
&= \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \sum_{k=0}^{n} \int_0^\infty \frac{C_n^k}{\Gamma(n+k)} s^{n+k-1} e^{-\tilde{\varphi}s} e^{-\nu s} ds \\
&= \int_0^\infty e^{-\nu s} \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n e^{-\tilde{\varphi}s} \sum_{k=0}^{n} \frac{C_n^k}{\Gamma(n+k)} s^{n+k-1} ds.
\end{aligned}$$

Obviously, the density function of S_n is the integrand except $e^{-\nu s}$. □

Then, the distribution of $Y_{t+1} = \theta *_g X_t$ is given by the following proposition.

Proposition 4.6. *The probability mass function of $Y_{t+1} = \varphi *_g X_t$ as well as its derivatives are given by*

$$\begin{aligned}
f_\varphi(y,n) &= P(Y_{t+1}=y|X_t=n) = \left(\frac{\tilde{\varphi}^2}{1+\tilde{\varphi}}\right)^n \sum_{k=0}^{n} C_n^k C_{n+k+y-1}^y (1+\tilde{\varphi})^{-(n+k+y)} \\
\frac{\partial f_\varphi(y,n)}{\partial \tilde{\varphi}} &= n\left(\frac{2}{\tilde{\varphi}} - \frac{1}{1+\tilde{\varphi}}\right) f_\varphi(y,n) - (y+1) f_\varphi(y+1,n) \\
\frac{\partial^2 f_\varphi(y,n)}{\partial \tilde{\varphi}^2} &= \left(n^2\left(\frac{2}{\tilde{\varphi}} - \frac{1}{1+\tilde{\varphi}}\right)^2 - n\left(\frac{2}{\tilde{\varphi}^2} - \frac{1}{(1+\tilde{\varphi})^2}\right)\right) f_\varphi(y,n) \\
&\quad - 2n(y+1)\left(\frac{2}{\tilde{\varphi}} - \frac{1}{1+\tilde{\varphi}}\right) f_\varphi(y+1,n) + (y+1)(y+2) f_\varphi(y+2,n) \\
\frac{\partial f_\varphi(y,n)}{\partial \varphi} &= \frac{\partial f_\varphi(y,n)}{\partial \tilde{\varphi}} \frac{\partial \tilde{\varphi}}{\partial \varphi}, \quad \frac{\partial^2 f_\varphi(y,n)}{\partial \varphi^2} = \frac{\partial^2 f_\varphi(y,n)}{\partial \tilde{\varphi}^2}\left(\frac{\partial \tilde{\varphi}}{\partial \varphi}\right)^2 + \frac{\partial f_\varphi(y,n)}{\partial \tilde{\varphi}} \frac{\partial^2 \tilde{\varphi}}{\partial \varphi^2},
\end{aligned} \tag{4.20}$$

where

$$\frac{\partial\tilde{\varphi}}{\partial\varphi} = -\frac{1}{2\varphi} + \frac{\varphi+3}{2\varphi\Delta} - \frac{1-\varphi+\Delta}{2\varphi^2}$$
$$\frac{\partial^2\tilde{\varphi}}{\partial\varphi^2} = \frac{1}{\varphi^2} + \frac{1}{2\varphi\Delta} + \frac{1-\varphi+\Delta}{\varphi^3} - \frac{(\varphi+3)^2}{2\varphi\Delta^3} - \frac{\varphi+3}{\varphi^2\Delta}$$

Proof.

$$\begin{aligned}
P(Y_{t+1}=y|X_t=n) &= \mathbb{E}\left[\frac{e^{-\sum_{i=1}^n\theta_i}(\sum_{i=1}^n\theta_i)^y}{y!}\right] \\
&= \int_0^\infty \frac{e^{-s}s^y}{y!}\left(\frac{\varphi^2}{1+\varphi}\right)^n e^{-\varphi s}\sum_{k=0}^n\frac{C_n^k}{\Gamma(n+k)}s^{n+k-1}ds \\
&= \left(\frac{\varphi^2}{1+\varphi}\right)^n\sum_{k=0}^n C_n^k\frac{\Gamma(n+k+y)}{\Gamma(n+k)\Gamma(y+1)}(1+\varphi)^{-(n+k+y)} \\
&= \left(\frac{\varphi^2}{1+\varphi}\right)^n\sum_{k=0}^n C_n^k C_{n+k+y-1}^y(1+\varphi)^{-(n+k+y)}
\end{aligned}$$

□

4.5 Maximum likelihood estimation and its asymptotic property

In general, the transition probability can be written down explicitly as

$$\begin{aligned}
P(X_{t+1}=i|X_t=j) &= \sum_{m=0}^{\min(i,j)} C_j^m p_1^m(1-p_1)^{j-m}P(Y_{t+1}+Z_{t+1}=i-m) \\
&= \sum_{m=0}^{\min(i,j)}\sum_{x=0}^{i-m} F_{p_1}(m,j)f_\varphi(x,j)F_{\lambda_1}(i-m-x) \\
F_{p_1}(m,j) &= C_j^m p_1^m(1-p_1)^{j-m} \\
f_\varphi(x,j) &= \int_{\mathbb{R}^+}\frac{e^{-s}s^x}{x!}g_j(s|\varphi)ds \\
F_{\lambda_1}(i-m-x) &= \frac{e^{-\lambda_1}\lambda_1^{i-m-x}}{(i-m-x)!}.
\end{aligned} \tag{4.21}$$

The log likelihood function is simply $\ell(p_1,\varphi,\alpha) = \sum_{t=0}^{n-1}\log P(X_{t+1}|X_t)$.

Proposition 4.7. *Suppose we have a random sample* $\{X_1, X_2, \ldots, X_n\}$. *Let* $\mathbf{p} =$

$(p_1, \varphi, \lambda_1)$ denote the parameters vector for the stationary BMP INAR(1) model. The maximum likelihood estimator $\hat{\mathbf{p}}$ has the following asymptotic distribution:

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \sim N(0, \mathbf{I}^{-1}), \tag{4.22}$$

where

$$\mathbf{H} = \begin{Bmatrix} \ell_{p_1 p_1} & \ell_{p_1 \varphi} & \ell_{p_1 \lambda_1} \\ \ell_{\varphi p_1} & \ell_{\varphi\varphi} & \ell_{\varphi\lambda_1} \\ \ell_{\lambda_1 p_1} & \ell_{p_1\varphi} & \ell_{\lambda_1\lambda_1} \end{Bmatrix} \quad \mathbf{I} = -\mathbb{E}[\mathbf{H}] \tag{4.23}$$

$$\begin{aligned}
\frac{\partial P(X_{t+1}|X_t)}{\partial p_1} &= \sum_{m=0}^{\min(X_{t+1},X_t)} \sum_{x=0}^{X_{t+1}-m} \frac{\partial F_{p_1}(m, X_t)}{\partial p_1} f_\varphi(x, X_t) F_{\lambda_1}(X_{t+1} - m - x) \\
\frac{\partial^2 P(X_{t+1}|X_t)}{\partial (p_1)^2} &= \sum_{m=0}^{\min(X_{t+1},X_t)} \sum_{x=0}^{X_{t+1}-m} \frac{\partial^2 F_{p_1}(m, X_t)}{\partial (p_1)^2} f_\varphi(x, X_t) F_{\lambda_1}(X_{t+1} - m - x) \\
\frac{\partial^2 P(X_{t+1}|X_t)}{\partial p_1 \partial \varphi} &= \sum_{m=0}^{\min(X_{t+1},X_t)} \sum_{x=0}^{X_{t+1}-m} \frac{\partial F_{p_1}(m, X_t)}{\partial p_1} \frac{\partial f_\varphi(x, X_t)}{\partial \varphi} F_{\lambda_1}(X_{t+1} - m - x) \\
\ell_{xy} &= \sum_{t=0}^{T-1} \frac{\partial^2 P(X_{t+1}|X_t)}{\partial x \partial y} \frac{1}{P(X_{t+1}|X_t)} - \frac{\partial P(X_{t+1}|X_t)}{\partial x} \frac{\partial P(X_{t+1}|X_t)}{\partial y} \frac{1}{P(X_{t+1}|X_t)^2},
\end{aligned} \tag{4.24}$$

where $x, y \in \{p_1, \varphi, \lambda_1\}$. The first and second derivatives of each distribution function is given by

$$\begin{aligned}
\frac{\partial F_{p_1}(m, X_t)}{\partial p_1} &= \frac{m - p_1 X_t}{p_1(1-p_1)} F_{p_1}(m, X_t) \\
\frac{\partial f_\varphi(m, X_t)}{\partial \varphi} &= \frac{\partial}{\partial \varphi} \int_{\mathbb{R}^+} \frac{e^{-s} s^x}{x!} g_{X_t}(s|\varphi) ds \\
\frac{\partial F_{\lambda_1}(m)}{\partial \lambda_1} &= \left(\frac{m}{\lambda_1} - 1 \right) F_{\lambda_1}(m)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 F_{p_1}(m, X_t)}{\partial (p_1)^2} &= \left(\frac{m(m - 1 - (X_t - 1)p_1)}{p_1^2(1-p_1)} - \frac{(X_t - m)(m - (X_t - 1)p_1)}{p_1(1-p_1)^2} \right) F_{p_1}(m, X_t) \\
\frac{\partial^2 f_\varphi(m, X_t)}{\partial \varphi^2} &= \frac{\partial^2}{\partial \varphi^2} \int_{\mathbb{R}^+} \frac{e^{-s} s^x}{x!} g_{X_t}(s|\varphi) ds \\
\frac{\partial^2 F_{\lambda_1}(x)}{\partial (\lambda_1)^2} &= \left(1 - \frac{2x}{\lambda_1} + \frac{x(x-1)}{\lambda_1^2} \right) F_{\lambda_1}(x)
\end{aligned}$$

Proof. From proposition 4.3, we know that the X_n is stationary and ergodic and its stationary distribution is characterized by the p.g.f $\Phi_x(\theta) = \prod_{i=0}^{\infty} \Phi_z(f_i(\theta))$. Then score functions and information matrix $\mathbf{I}$ are also stationary and ergodic. Then the proof for asymptotic normality is similar to the proof of theorem 4 in Appendix A of Bu et al. (2008) □

The expectation of information matrix $\mathbf{I}$ can be calculated numerically by finding out unconditional distribution $P(X_t)$ and joint distribution $P(X_{t-1}, X_t)$. However, this would be computational intensive when the sample size n is large. In practice, since the process X_t is stationary and ergodic, $\mathbf{I} \approx -\mathbf{H}$ when n is large.

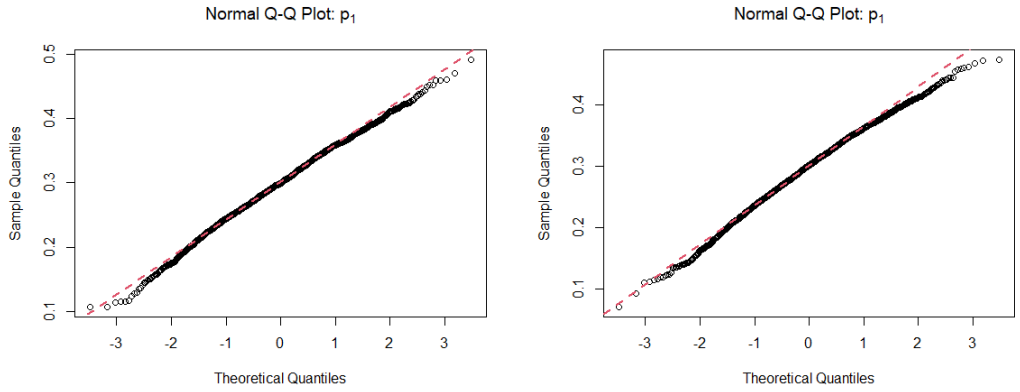
To verify the asymptotic normality of the maximum likelihood estimators, we conduct a Monte Carlo experiment. This experiment is based on 2000 replications. For each replication, a time series of BMP-INAR(1) with chosen mixing density, either Exponential or Lindley , of size $n = 100, 200, \ldots, 500$ is generated. The parameters are set as $p_1 = \varphi = 0.3, \lambda_1 = 2$ for both mixing densities and they are estimated via the maximum likelihood method. The biases and standard errors of the estimated parameters are shown in tables 1 and 2. We observe that the biases of the estimators are either reasonably small or decreasing with respect to the sample size n. And it is clear that the standard error is also decreasing with respect to n. Finally, in order to graphically inspect the distribution of estimators, normal quantile-quantile plots are provided below.

Table 4.1: The bias of Maximum likelihood estimators of BMP-INAR(1) model with respect to different sample size n

	Bias($\hat{\mathbf{p}}$)	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
Exponential	p_1	0.0022	-0.0021	0.0019	-0.0003	-0.0003
	ϕ	-0.0284	-0.0104	-0.0110	-0.0072	-0.0059
	λ_1	0.1089	0.0526	0.0384	0.0366	0.0279
Lindley	p_1	-0.0008	0.0004	-0.0015	-0.0020	-0.0011
	ϕ	-0.0209	-0.0143	-0.0085	-0.0050	-0.0039
	λ_1	0.0387	0.0227	0.0141	0.0144	0.0101

Table 4.2: The standard error of Maximum likelihood estimators of BMP-INAR(1) model with respect to different sample size n

	S.E.$(\hat{\mathbf{p}})$	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
Exponential	p_1	0.1303	0.0965	0.0752	0.0663	0.0576
	ϕ	0.1384	0.0970	0.0783	0.0670	0.0581
	λ_1	0.3982	0.2858	0.2276	0.2012	0.1764
Lindley	p_1	0.1319	0.0991	0.0854	0.0711	0.0630
	ϕ	0.1432	0.1054	0.0880	0.0729	0.0661
	λ_1	0.2050	0.1515	0.1166	0.0999	0.0911



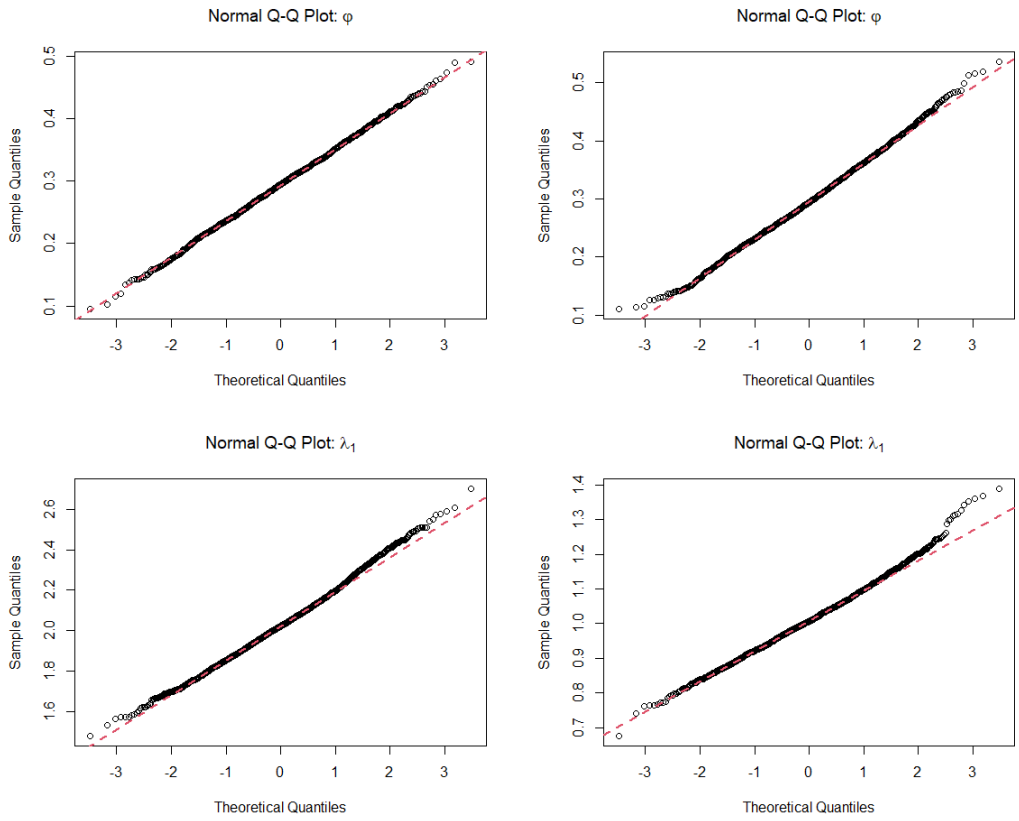


Figure 4.1: Quantile-Quantile plots for maximum likelihood estimators of BMP-INAR(1) model. The left panel shows plots for the Exponential mixing density, while the right panel depicts the plots for the Lindley mixing density.

4.6 Real data example: iceberg order data

The iceberg order counts concern the Deutsche Telekom shares traded in the XETRA system of Deutsche Börse, and the concrete time series gives the number of iceberg orders (for the ask side) per 20 min for 32 consecutive trading days in the first quarter of 2004. The special feature of iceberg orders is that only a small part of the order (tip of the iceberg) is visible in the order book and the main part of the order is hidden. For detail description, please see the Jung and Tremayne (2011) and Frey and Sandås (2009). This dataset is also analysed in Weiß (2015), where the Extended Poisson INAR(1) is applied to fit the data.

A table of descriptive statistics, a time series, as well as the ACF and PACF plots are shown below. The variance of the iceberg count is higher than its mean, which indicates the data is overdispersed. The level of dispersion is described by the Fisher index of dispersion FI > 1. Evidence of the applicability of a first order autoregressive model is indicated by the empirical ACF and PACF graphs. They

illustrate a clear decay for ACF and cut-off at lag =1 for PACF.

Table 4.3: Descriptive statistics of iceberg count

minimum	maximum	median	mean	variance	FI
0	9	1	1.407	2.184	1.552

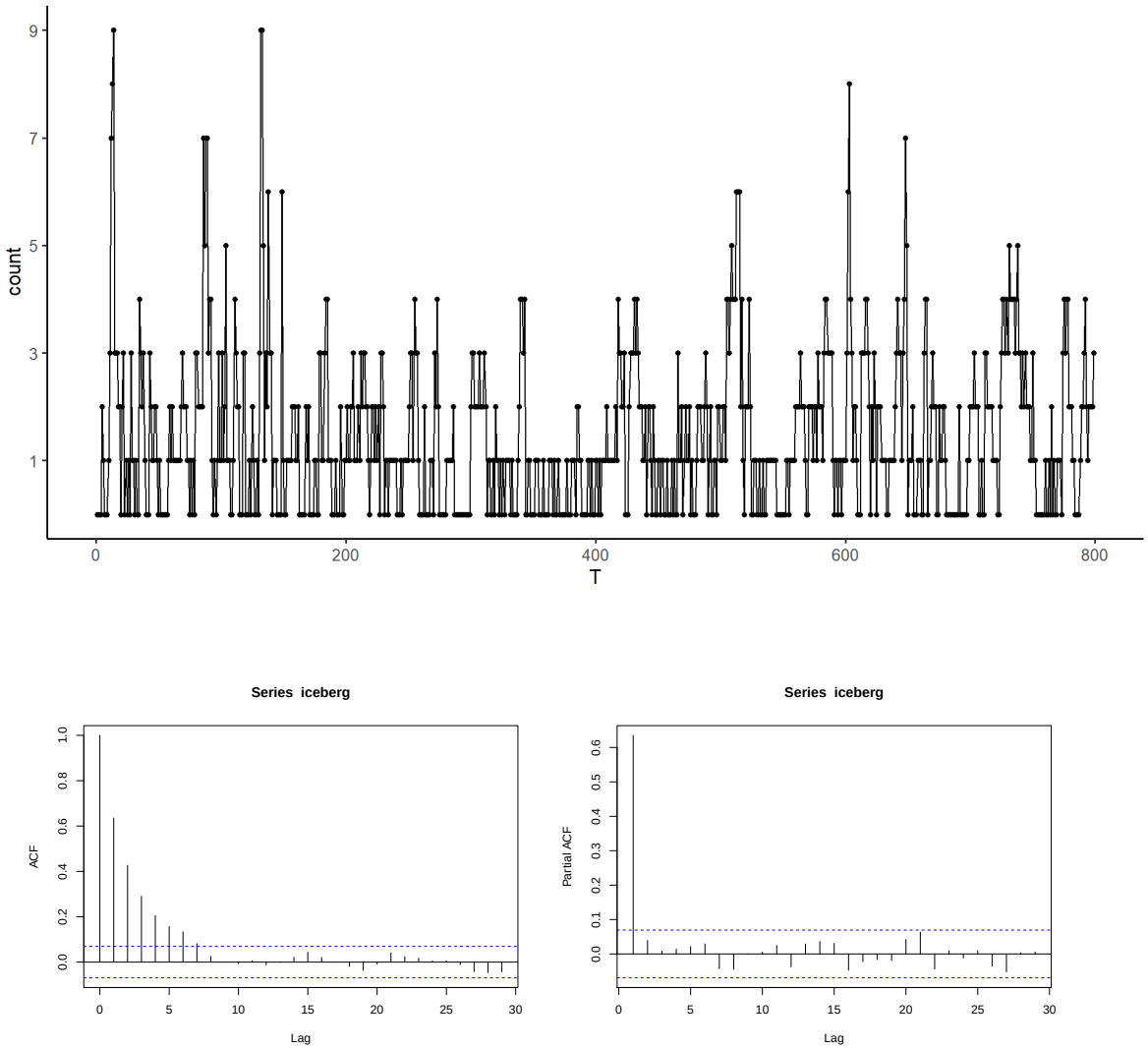


Figure 4.2: Time series plot of iceberg data and its emprical ACF PACF plots. The blue dash lines are the 95% confident bands by assuming the series to be a white noise process

The following table records the estimated parameters through the maximum likelihood method. The likelihood function is constructed as in 4.21 with different $f_\varphi(x, j)$ (mixed by Exponential or Lindley). It is then maximised through 'optim' in R with 'method = BFGS' (quasi-Newton method) while the standard deviations of MLEs are calculated through inverting the negative observed information matrix in proposition 4.7 based on MLEs. To access the goodness of fit, we adopt the information criteria AIC and BIC as well as the (standardized) Pearson residuals. If the model

is correctly specified, Pearson residuals for BMP-INAR(1) are expected to have a mean and variance close to 0 and 1 respectively, with no significant autocorrelation. The Pearson residuals are calculated by the following formula:

$$e_t = \frac{x_t - \mathbb{E}[X_t|x_{t-1}]}{\sqrt{Var(X_t|x_{t-1})}}, \tag{4.25}$$

where x_t denotes the observed value.

Table 4.4: The results for the BMP INAR(1) model mixed by different density functions. The results of Dirac delta case are from Table 2 of Weiß (2015). The estimated standard deviations for all models are in brackets.

Mixing density	$\hat{p_1}$	$\hat{\varphi}$	$\hat{\lambda_1}$	AIC	BIC	Pearson residuals mean	Pearson residuals variance	$\hat{\text{FI}}_x$
Dirac delta	0.410	0.188	0.567	2212	2226	-0.001	1.159	1.295
	(0.058)	(0.059)	(0.040)					
Exponential	0.434	0.167	0.563	2208	2222	-0.002	1.154	1.315
	(0.044)	(0.044)	(0.040)					
Lindley	0.434	0.167	0.563	2208	2222	-0.002	1.154	1.314
	(0.043)	(0.043)	(0.040)					

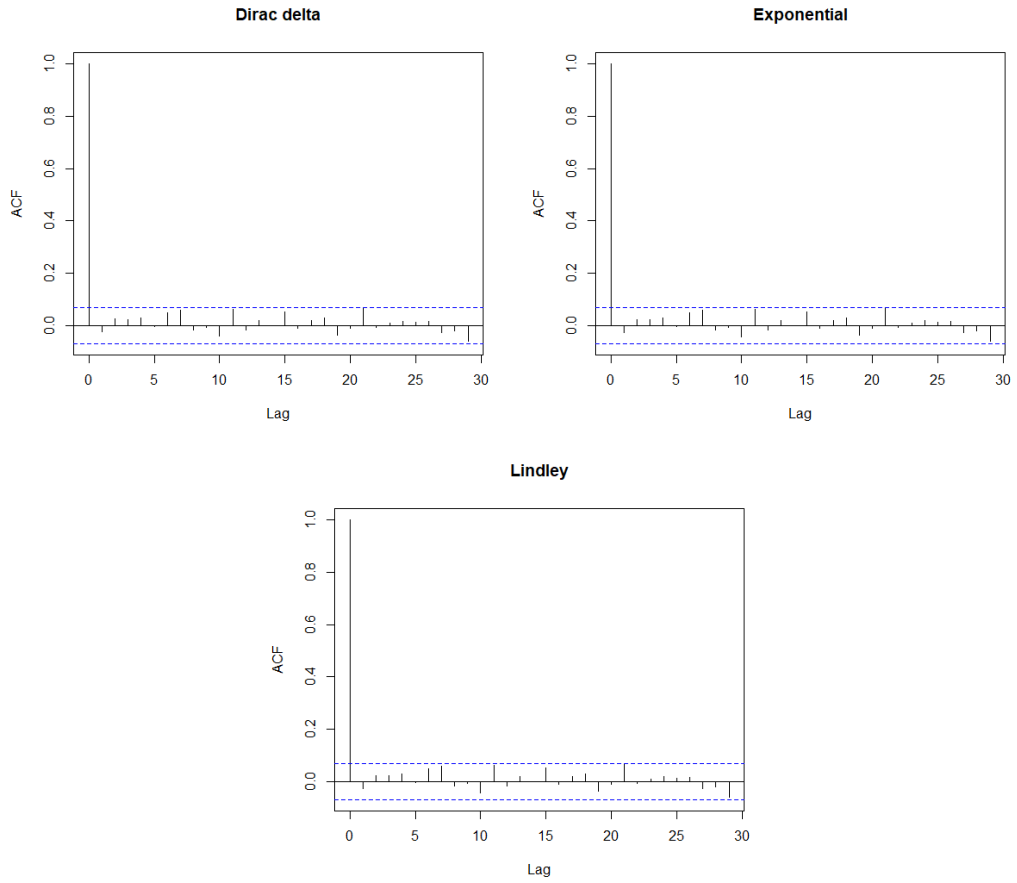


Figure 4.3: Autocorrelation of Standardized Pearson residuals for three different mixing densities

The results shown in Table 2 and the ACF plots of the Pearson residuals indicate that the BMP-INAR(1) models are appropriate for fitting the iceberg data. The estimated parameters are significantly different from 0, which is indicated by their estimated standard deviation. Compared to the Dirac delta case, which is actually the Extended Poisson INAR(1) of Weiß (2015), the other two cases do show some improvement with smaller AIC, BIC values and larger fitted Fisher index of dispersion $\hat{\mathrm{FI}}_x$ which, however, is slightly smaller than the empirical FI. On the other hand, it seems that there is little difference between the other two cases as they have very similar AIC and BIC values. This is due to the fact that the value of $\hat{\varphi}$ is identical for both densities. Finally, it should be noted that the variance of the Pearson residuals is visibly larger than 1. As it was previously mentioned, the exponential and Lindley mixing densities were considered for expository purposes. Therefore, since the proposed family of BMP INAR(1) models is quite general, another mixing distribution could potentially more efficiently capture the observed dispersion structure for this data.

Overall, the mixed Poison component in the BMP INAR(1) model efficiently captures the overdispersion in this type of financial data.

4.7 Concluding remarks

The BMP INAR(1) is an extension of the classical Poisson INAR(1) model obtained by adding an additional mixed Poisson component and hence it can capture the level of overdispersion coming from the data. The exponential family is a desired choice for the mixing density due to its 'additivity' property. The choice of the mixing density can control the dispersion level to some extent, although the BMP INAR(1) X_t is always overdispersed in general. Furthermore, due to its simplicity, X_t is actually a Markov chain and the maximum likelihood estimation method can be applied easily. The real data analysis shows that BMP INAR(1) can be a potential choice for modelling financial count data which exhibit standard AR(1) structure and overdispersion.

CHAPTER 5

Paper B. Multivariate Mixed Poisson Generalized Inverse Gaussian INAR(1) Regression

Abstract

In this paper, we present a novel family of multivariate mixed Poisson-Generalized Inverse Gaussian INAR(1), MMPGIG-INAR(1), regression models for modelling time series of overdispersed count response variables in a versatile manner. The statistical properties associated with the proposed family of models are discussed and we derive the joint distribution of innovations across all the sequences. Finally, for illustrative purposes different members of the MMPGIG-INAR(1) class are fitted to Local Government Property Insurance Fund data fromthe state of Wisconsin via maximum likelihood estimation.

5.1 Introduction

In recent years, there has been a growing interest in modelling integer-valued time series of univariate and multivariate count data in a plethora of different scientific fields such as sociology, econometrics, manufacturing, engineering, agriculture, biology, biometrics, genetics, medicine, sports, marketing, and insurance. In particular, regarding the univariate case Al-Osh and Alzaid (1987) and McKenzie (1985) were the first to consider an INAR(1) model based on the so-called binomial thinning

operator. Subsequently, many articles focused on extending this setup by applying different thinning operators or by varying the distribution of innovations. For more details, the interested reader can refer to Weiß (2018) Davis et al. (2016a), Scotto et al. (2015), Weiß (2008b) among many more. The INAR(1) model with Poisson marginal distribution (Poisson INAR(1)) has been the most popular choice due to the simplicity of its log-likelihood function that implies that the formality of parameter estimation via maximum likelihood (ML) estimation is straightforward. Also, Freeland and McCabe (2004) considered an extension of the model by allowing for regression specifications on the mean of the Poisson innovation as well as parameter of binomial thinning operator. On the other hand, the literature which focuses on the multivariate case is less developed. In particular, Latour (1997) introduced a multivariate GINAR(p) model with a generalized thinning operator. Karlis and Pedeli (2013) and Pedeli and Karlis (2011, 2013a,b) focused on the diagonal case under which the thinning operators do not introduce cross correlation among different counts. In this case, the dependence structure introduced by innovations. Additionally, Ristić et al. (2012), Popović (2016), Popović et al. (2016) and Nastić et al. (2016) constructed multivariate INAR distributions with cross correlations among counts and random coefficients thinning. Finally, Karlis and Pedeli (2013) extended the setup of the previous articles by allowing for negative cross correlation via a copula-based approach for modelling the innovations.

In this paper, we extend the model proposed by Pedeli and Karlis (2011) by introducing the multivariate mixed Poisson-Generalized Inverse Gaussian INAR(1), MMPGIG-INAR(1), regression model for multivariate count time series data. The MMPGIG-INAR(1) is a general three parameter distribution family of INAR(1) models driven by mixed Poisson regression innovations where the mixing densities are chosen from the Generalized Inverse Gaussian class of distributions. Thus, the proposed modelling framework can provide the appropriate level of flexibility for modelling positive correlations of different magnitudes among time series of different types of overdispersed count response variables. In particular, depending on the values taken by the shape parameter, the MMPGIG-INAR(1) family includes many members, such as the mixed Poisson-Inverse Gaussian (PIG), as special cases and several others as limiting cases, such as the Negative Binomial, or Poisson-Gamma, the Poisson-Inverse Gamma (PIGA), the Poisson-Inverse Exponential, the Poisson-Inverse Chi Squared and the Poisson-Scaled Inverse Chi Squared distributions. Therefore, it can accommodate different levels of overdispersion depending on the chosen parametric form of the mixing density. Furthermore, the MMPGIG-INAR(1) family of models is constructed by assuming that the probability mass function (pmf) of the MMPGIG innovations is parameterized in terms of the mean

parameter which results in a more orthogonal parameterization that facilitates maximum likelihood (ML) estimation when regression specifications are allowed for the mean parameters of the MMPGIG-INAR(1) regression model. For expository purposes, we derive the joint probability mass functions and the derivatives of several special cases of the MMPGIG-INAR(1) family which are used as innovations. These models are fitted to time series of claim count data from the Local Government Property Insurance Fund (LGPIF) data in the state of Wisconsin. At this point it is worth noting that modelling the correlation between different types of claims from the same and/or different types of coverage it is very important from a practical business standpoint. Many articles have been devoted to this topic, see for example, Bermúdez and Karlis (2011), Bermúdez and Karlis (2012), Shi and Valdez (2014a), Shi and Valdez (2014b), Abdallah et al. (2016), Bermúdez and Karlis (2017), Pechon et al. (2018), Pechon et al. (2019), Bolancé and Vernic (2019), Denuit et al. (2019), Fung et al. (2019), Bolancé et al. (2020), Pechon et al. (2021), Jeong and Dey (2021), Gómez-Déniz and Calderín-Ojeda (2021), Tzougas and di Cerchiara (2023) and Tzougas and di Cerchiara (2021).

However, with the exception of very few articles, such as Bermúdez et al. (2018) and Bermúdez and Karlis (2021), the construction of bivariate INAR(1) models which can capture the serial correlation between the observations of the same policyholder over time and the correlation between different claim types remains a largely uncharted territory. This is an additional contribution of this study.

The rest of the paper proceeds as follows. Section 2 presents the derivation of the MMPGIG-INAR(1) model. Statistical properties of the MMPGIG innovations are discussed in Section 3. In Section 4, we present a description of the alternative special cases of the MMPGIG-INAR(1) family. Section 5 discusses the parameter estimation for these models based on the maximum likelihood method and integer-valued prediction. Section 6 contains our empirical analysis for the LGPIF data set. Finally, concluding remarks are given in Section 7.

5.2 Generalized Setting

Let $\mathbf{X}$ and $\mathbf{R}$ be non-negative integer-valued random vectors in $\mathbb{R}^m$. Let $\mathbf{P}$ be a diagonal matrix in $\mathbb{R}^{m\times m}$ with elements $p_i \in (0,1)$. The multivariate Poisson-

Generalized Inverse Gaussian INAR(1) is defined as

$$\mathbf{X}_t = \mathbf{P} \circ \mathbf{X}_{t-1} + \mathbf{R}_t = \begin{bmatrix} p_1 & 0 & \dots & 0 & 0 \\ 0 & p_2 & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & & p_m \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ \vdots \\ X_{m,t-1} \end{bmatrix} + \begin{bmatrix} R_{1,t} \\ R_{2,t} \\ \vdots \\ R_{m,t} \end{bmatrix} \tag{5.1}$$

where the thinning operator $\circ$ is the widely used binomial thinning operator such that $p_i \circ X_{i,t} = \sum_{k=1}^{X_{i,t}} U_k$ where U_k are independent identically distributed Bernoulli random variables with success probability p_i, i.e. $\mathcal{P}(U_k = 1) = p_i$. Hence $p_i \circ X_{i,t}$ is binomially distributed with size $X_{i,t}$ and success probability p_i. Then the distribution function $f_{p_i}(x, X_{i,t})$ can be easily written down as

$$f_{p_i}(x, X_{i,t}) = \binom{X_{i,t}}{x} p_i^x (1-p_i)^{X_{i,t}-x} \tag{5.2}$$

Note that given $X_{i,t}, X_{j,t}$ $i \neq j$, $p_i \circ X_{i,t}$ and $p_j \circ X_{j,t}$, are independent of each other. To adapt the heteroscedasticity arising from the data, $\{R_{i,t}\}_{i=1,\dots,m}$ are mixed Poisson random variables $Po(\theta_t \lambda_{i,t})$ with the random effect θ_t. The rate $\lambda_{i,t}$ is characterized by its observed covariate $z_{i,t} \in \mathbb{R}^{a_i \times 1}$ for some positive integer a_i and they are connected through a log link function such that $\log(\lambda_{i,t}) = z_{i,t}^T \beta_i$ where $\beta_i \in \mathbb{R}^{a_i \times 1}$. Furthermore, $\{R_{i,t}\}_{i=1,\dots,m}$ share the same random effect θ_t with mixing distribution $G(\theta)$, which means the dependent structure among $X_{i,t}$ can be controlled by the choice of distribution and its corresponding size of parameters. The joint distribution of $\mathbf{R}_t$ is

$$\begin{aligned} f_\phi(\mathbf{k}, t) &= \mathcal{P}(R_{1,t} = k_1, \dots, R_{m,t} = k_m) \\ &= \mathbb{E}\left[\mathcal{P}(R_{1,t} = k_1, \dots, R_{m,t} = k_m | \theta_t)\right] \\ &= \prod_{j=1}^{m} \frac{\lambda_{j,t}^{k_j}}{k_j!} \int_0^\infty e^{-\theta \sum_{i=1}^m \lambda_{i,t}} \theta^{\sum_{i=1}^m k_i} dG(\theta) \end{aligned} \tag{5.3}$$

We let θ_t be a continuous random variable from the Generalized Inverse Gaussian distribution with density function $g(\theta)$

$$g(\theta) = \frac{(\psi/\chi)^{\frac{\nu}{2}}}{2K_\nu(\sqrt{\psi\chi})} \theta^{\nu-1} \exp\left\{-\frac{1}{2}\left(\psi\theta + \frac{\chi}{\theta}\right)\right\}, \tag{5.4}$$

where $-\infty \leqslant \nu \leqslant \infty, \psi > 0, \chi > 0$ and $K_\nu(\omega)$ is the modified Bessel function of the third kind of order ν and argument ω such that

$$K_\nu(\omega) = \int_0^\infty z^{\nu-1} \exp\left\{-\frac{1}{2}\omega(z+\frac{1}{z})\right\} dz$$

The Generalized Inverse Gaussian distribution is a widely used family. For example, it includes the Inverse Gaussian as special case and the Gamma and Inverse Gamma as limiting cases. To avoid identification problems for mixed Poisson regression random variable $\mathbf{R}_t$, the mean of θ_t is restricted to one, i.e. $\mathbb{E}[\theta_t] = 1$, and all the parameters ν, ψ, χ will be either fixed or a function of another parameter ϕ. With these two constraints, there is only one parameter that is free to vary. (e.g. for Inverse Gaussian distribution, $\nu = -\frac{1}{2}$ and $\psi = \chi = \phi$). The joint distribution of $\mathbf{R}_t$ becomes an MPGIG distribution

$$\begin{aligned} f_\phi(\mathbf{k}, t) &= \frac{(\psi/\chi)^{\frac{\nu}{2}}}{2K_\nu(\sqrt{\psi\chi})} \prod_{j=1}^m \frac{\lambda_{j,t}^{k_j}}{k_j!} \int_0^\infty e^{-\theta\sum_{i=1}^m \lambda_{i,t}} \theta^{\sum_{i=1}^m k_i} \theta^{\nu-1} \exp\left\{-\frac{1}{2}\left(\psi\theta + \frac{\chi}{\theta}\right)\right\} d\theta \\ &= \frac{(\psi/\chi)^{\frac{\nu}{2}}}{(\Delta/\chi)^{\frac{\nu+\sum_i k_i}{2}}} \frac{K_{\nu+\sum_i k_i}(\sqrt{\Delta\chi})}{K_\nu(\sqrt{\psi\chi})} \prod_{j=1}^m \frac{\lambda_{j,t}^{k_j}}{k_j!}, \end{aligned} \tag{5.5}$$

where $\Delta = \psi + 2\sum_{i=1}^m \lambda_{i,t}$. In section 5, we will discuss in detail the distribution function $f_\phi(\mathbf{k}, t)$ for some special cases. Finally, it should be noted that several articles discuss multivariate versions of MPGIG distribution and/or the MPIG distribution which is a special case for $\nu = -0.5$, see, for instance, Barndorff-Nielsen et al. (1992), Ghitany et al. (2012) Amalia et al. (2017), Mardalena et al. (2020), Tzougas and di Cerchiara (2023) and Mardalena et al. (2021). However, this is the first time that the MMPGIG-INAR(1) distribution family of INAR(1) models driven by mixed Poisson regression innovations are considered for modelling time series of count response variables.

5.3 Properties of innovations $\mathbf{R}_t$

Proposition 5.1. *(The moments of $\mathbf{R}_t$) The mean, variance of $R_{i,t}$ and covariance between $R_{i,t}, R_{j,t}, i \neq j$ are given by*

$$\begin{aligned}
\mathbb{E}[R_{i,t}] &= \mathbb{E}[\mathbb{E}[R_{i,t}|\theta_t]] = \lambda_{i,t} \\
Var(R_{i,t}) &= Var(\mathbb{E}[R_{i,t}|\theta_t]) + \mathbb{E}[Var(R_{i,t}|\theta_t)] \\
&= \sigma_\theta^2 \lambda_{i,t}^2 + \lambda_{i,t} \\
Cov(R_{i,t}, R_{j,t}) &= Cov(\mathbb{E}[R_{i,t}|\theta_t], \mathbb{E}[R_{j,t}|\theta_t]) + \mathbb{E}[Cov(R_{i,t}, R_{j,t}|\theta_t)] \\
&= \lambda_{i,t}\lambda_{j,t}\sigma_\theta^2
\end{aligned} \tag{5.6}$$

where σ_θ^2 is the variance for the random effect θ_t and $i, j = 1, \ldots, m$.

Proposition 5.2. *(Marginal property) The joint distribution function $f_\phi(\mathbf{k}, t)$ is closed to marginalization, i.e. the marginal distribution for $R_{i,t}$ is given by $f_\phi(k_i, t)$ such that*

$$\begin{aligned}
f_\phi(k_i, t) &= \int_0^\infty \frac{\lambda_{i,t}^{k_i}}{k_i!} \theta^{k_i} e^{-\lambda_{i,t}\theta} dG(\theta) \\
&= \frac{(\psi/\chi)^{\frac{\nu}{2}}}{((\psi + \lambda_{i,t})/\chi)^{\frac{\nu+k_i}{2}}} \frac{K_{\nu+k_i}(\sqrt{(\psi + \lambda_{i,t})\chi})}{K_\nu(\sqrt{\psi\chi})} \frac{\lambda_{i,t}^{k_i}}{k_i!}
\end{aligned} \tag{5.7}$$

which is a univariate mixed Poisson regression random variable. In general, this result is valid for any m'-variate mixed Poisson regression random variable with $m' < m$

Proof. We will show the result for univariate case. The m'-variate case can be derived similarly by reducing the number of following sum to $m - m'$

$$\begin{aligned}
f_\phi(k_i, t) &= \sum_{k_1=0} \cdots \sum_{k_{i-1}=0} \sum_{k_{i+1}=0} \cdots \sum_{k_m=0} f_\phi(\mathbf{k}, t) \\
&= \int_0^\infty \prod_{j \neq i} \left(\sum_{k_j=0}^\infty \frac{e^{-\theta\lambda_{j,t}} (\theta\lambda_{j,t})^{k_j}}{k_j!} \right) \frac{e^{-\theta\lambda_{i,t}} (\theta\lambda_{i,t})^{k_i}}{k_i!} dG(\theta) \\
&= \int_0^\infty \frac{\lambda_{i,t}^{k_i}}{k_i!} \theta^{k_i} e^{-\lambda_{i,t}\theta} dG(\theta)
\end{aligned}$$

□ □

The marginalization property can enable, for example insurers, to easily price those policyholders who only engage in some but not all lines of business. The last property is about the identifiability of $\mathbf{R}_t$, which will ensure the uniqueness of the model.

Proposition 5.3. *(Identifiability of joint distribution $\mathbf{R}_t$) Assume that the covariate space $\mathbf{z}_t = (z_{1,t}, \ldots, z_{m,t})$ is of full rank. Denote the parameter set $\Theta_R = \{\beta_i, \phi | i = 1, \ldots, m\}$ and $\tilde{\Theta}_R = \{\tilde{\beta}_i, \tilde{\phi} | i = 1, \ldots, m\}$, the joint distribution $f_\phi(\mathbf{k}, t)$ is identifiable such that*

$$f_\phi(\mathbf{k}, t) = f_{\tilde{\phi}}(\mathbf{k}, t)$$

if and only if $\Theta_R = \tilde{\Theta}_R$.

Proof. With the assumption that the covariate $\mathbf{z}$ is of full rank and the log-link function is monotonic such that $\log(\lambda_{i,t}) = z_{i,t}^T \beta_i$, it is obvious that the identification problem for the mixed Poisson regression random variable $\mathbf{R}_t$ reduces to identification for mixed Poisson random variable (without regression), which means the set of parameter can be re-parametrized as $\Theta_R^* = \{\lambda_{i,t}, \phi | i = 1, \ldots, m\}$ and $\tilde{\Theta}_R^* = \{\tilde{\lambda}_{i,t}, \tilde{\phi} | i = 1, \ldots, m\}$.
Then the 'if' statement is obvious since the same set of parameters will definitely lead to the same joint distribution function. For the 'only if' statement, to match two distribution functions, all the moments (mean,variance, covariance) must reconcile. From the moment properties above, matching the $\mathbb{E}[R_{i,t}]$ will lead to $\lambda_{i,t} = \tilde{\lambda}_{i,t}$. Likewise, given that the first moment is matched, only $\phi = \tilde{\phi}$ will lead to the same $Var(R_{i,t})$. Matching these moments already leads to $\Theta_R^* = \tilde{\Theta}_R^*$, then the covariance $Cov(R_{i,t}, R_{j,t})$ must match with each other. □

5.4 Model specification

The distributional properties of $\mathbf{X}_t$, in particular the correlation structure and 'tailedness' of the distribution, are mainly determined by the innovation $\mathbf{R}_t$, more specifically, the mixing density $g(\theta)$. On the other hand, the explicit form of the derivatives of $f_\phi(\mathbf{k}, t)$ can significantly accelerate the computational speed when performing estimation. Hence, the distribution function $f_\phi(\mathbf{k}, t)$ as well as its derivatives are derived for two limiting cases (Gamma, Inverse Gamma) and some other special cases (GIG with unit mean and different values of ν). Throughout this session, we define $S_t^\lambda = \sum_{i=1}^m \lambda_{i,t}$ and $S^k = \sum_{i=1}^m k_i$.

5.4.1 Mixing by Gamma distribution

If $\mathbf{R}_t$ is univariate, the resulting distribution is known as the negative binomial distribution and this result can be easily extended to the multivariate case which is called the multivariate negative binomial distribution (see e.g. Marshall and Olkin (1990) Boucher et al. (2008) Cheon et al. (2009)). The gamma density is obtained by letting $\nu = \phi, \psi = 2\phi$ and $\chi = 0$ in generalized Inverse Gaussian density in 5.4. The resulting mixing density has the following form:

$$g(\theta) = \frac{\phi^\phi}{\Gamma(\phi)} \theta^{\phi-1} e^{-\phi\theta} \tag{5.8}$$

with unit mean and variance $\frac{1}{\phi}$. Then the expectation 5.3 can be evaluated explicitly

$$\begin{aligned} f_\phi(\mathbf{k}, t) &= \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \mathbb{E}[e^{-(S_t^\lambda)\theta} \theta^{S^k}] \\ &= \frac{\Gamma(\phi + S^k)}{\Gamma(\phi) \prod_{i=1}^{m} \Gamma(k_i + 1)} \frac{\phi^\phi \prod_{i=1}^{m} \lambda_{i,t}^{k_i}}{(\phi + S_t^\lambda)^{\phi + S^k}} \end{aligned} \tag{5.9}$$

Proposition 5.4. *The derivatives of the distribution function $f_\phi(\mathbf{k}, t)$ with respect to $\Theta_R = \{\phi, \beta_i \mid i = 1, \ldots, m\}$ when $\theta_t \sim$ Gamma(ϕ, ϕ) are given by*

$$\begin{aligned} \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \phi} &= f_\phi(\mathbf{k}, t) \left(\sum_{n=1}^{S^k} \frac{1}{n + \phi - 1} + \log\left(\frac{\phi}{\phi + S_t^\lambda} \right) + \frac{\sum_{i=1}^{m} (\lambda_{i,t} - k_i)}{\phi + S_t^\lambda} \right) \\ \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \beta_i} &= f_\phi(\mathbf{k}, t) \left(\frac{k_i}{\lambda_{i,t}} - \frac{\phi + S^k}{\phi + S_t^\lambda} \right) \lambda_{i,t} z_{i,t} \ , \end{aligned} \tag{5.10}$$

where the sum $\sum_{n=1}^{S^k} \frac{1}{n+\phi-1} = 0$ when $S^k = 0$.

Proof. The derivatives $\frac{\partial f_\phi(\mathbf{k},t)}{\partial \beta_i}$ can be figured out easily except $\frac{\partial f_\phi(\mathbf{k},t)}{\partial \phi}$ which involves the gamma function. The derivative of the gamma function can be derived by utilizing the alternative Weierstrass's definition such that

$$\Gamma(z+1) = e^{-\gamma z} \prod_{n \geqslant 1} (1 + \frac{z}{n})^{-1} e^{\frac{z}{n}},$$

which is valid for all complex number z except non-positive integers and γ is Euler–Mascheroni constant. Then the derivative can be derived by differentiating its

log transform $\log\Gamma(z+1)$, which leads to the series expansion of digamma function

$$\Psi(z+1) = \frac{\Gamma'(z+1)}{\Gamma(z+1)} = -\gamma + \sum_{n\geqslant 1}\left(\frac{1}{n} - \frac{1}{n+z}\right)$$

Then the derivative $\frac{\partial f_\phi(\mathbf{k},t)}{\partial\phi}$ can be derived steps by steps. First let us simplify the expression of $f_\phi(\mathbf{k},t)$ such that

$$f_\phi(\mathbf{k},t) = c_1\frac{N(\phi)}{D(\phi)}$$
$$c_1 = \prod_{i=1}^{m}\frac{\lambda_{i,t}^{k_i}}{\Gamma(k_i+1)},\ N(\phi) = \Gamma(\phi+S^k)\phi^\phi,\ D(\phi) = \Gamma(\phi)\left(\phi + S_t^\lambda\right)^{\phi+S^k}$$

The derivative is then

$$\frac{\partial f_\phi(\mathbf{k},t)}{\partial\phi} = c_1\frac{N'(\phi)D(\phi) - N(\phi)D'(\phi)}{D^2(\phi)}$$
$$= c_1\frac{N(\phi)}{D(\phi)}\left(\sum_{n\geqslant 1}\left(\frac{1}{n+\phi-1} - \frac{1}{n+\phi+S^k-1}\right) + 1 + \log\phi - \log(\phi+S_t^\lambda) - \frac{\phi+S^k}{\phi+S_t^\lambda}\right)$$

□ □

5.4.2 Mixing by Inverse Gamma

The Inverse gamma distribution, which is another limiting case of generalized Inverse Gaussian distribution, is discussed in section 9.3 Johnson et al. (1995). Inverse gamma random variable has a relatively thicker right tail and a low probability in taking the values closed to 0. In this case, the density function $g(\theta)$ is obtained by letting $\psi = 0, \chi = 2\phi$ and $\nu = -\phi - 1$ such that

$$g(\theta) = \frac{\phi^{\phi+1}}{\Gamma(\phi+1)}\theta^{-\phi-2}e^{-\frac{\phi}{\theta}}, \tag{5.11}$$

with mean 1 and variance $\frac{1}{\phi-1}$ for $\phi > 1$. It is also called the reciprocal gamma distribution such that $\theta = 1/x$ where $x \sim \text{Gamma}(\phi+1,\phi)$. The distribution

function $f_\phi(\mathbf{k}, t)$ becomes

$$\begin{aligned} f_\phi(\mathbf{k}, t) &= \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \mathbb{E}\left[e^{-S_t^\lambda \theta_t} \theta_t^{S^k}\right] \\ &= \frac{2K_\nu(\omega)}{\Gamma(\phi+1)\prod_{i=1}^{m}\Gamma(k_i+1)} \frac{\phi^{\frac{\nu}{2}+\phi+1}\prod_{i=1}^{m}\lambda_{i,t}^{k_i}}{(S_t^\lambda)^{\frac{\nu}{2}}}, \end{aligned} \tag{5.12}$$

where $\nu = S^k - \phi - 1$ and $\omega = 2\sqrt{\phi S_t^\lambda}$. The derivatives of $f_\phi(\mathbf{k}, t)$ with respect to the parameter set $\Theta_R = \{\phi, \beta_i \mid i = 1, \ldots, m\}$ are given by

$$\begin{aligned} \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \phi} &= \left(\log\frac{\omega}{2} + \frac{S^k + \phi + 1}{2\phi} + \frac{\partial \log K_\nu(\omega)}{\partial \phi} - \Psi(\phi+1)\right) f_\phi(\mathbf{k}, t) \\ \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \beta_i} &= \left(\frac{k_i}{\lambda_{i,t}} f_\phi(\mathbf{k}, t) - \frac{k_i+1}{\lambda_{i,t}} f_\phi(\mathbf{k}+\mathbf{1}_i, \phi)\right) \lambda_{i,t} z_{i,t} \end{aligned} \tag{5.13}$$

In this case, numerical differentiation is applied to calculate $\frac{\partial \log K_\nu(\omega)}{\partial \phi}$ since the parameter ϕ appears both in the order ν and argument ω of the modified Bessel function $K_\nu(\omega)$.

5.4.3 Mixing by Generalized Inverse Gaussian

Likewise, if $\mathbf{R}_t$ is univariate, the distribution of $\mathbb{R}_t$ is known as the Poisson Generalized Inverse Gaussian distribution. To comply with constraints we made in section 2, the mixing density function has following form

$$g(\theta) = \frac{c^\nu}{2K_\nu(\phi)} \theta^{\nu-1} \exp\left\{-\frac{\phi}{2}\left(c\theta + \frac{1}{c\theta}\right)\right\} \tag{5.14}$$

with unit mean and variance $var(\theta_t) = \frac{1}{c^2} + \frac{2(\nu+1)}{c\phi} - 1$, where $c = \frac{K_{\nu+1}(\phi)}{K_\nu(\phi)}$, $\phi > 0$ and $\nu \in \mathbb{R}$. Then the distribution function $f_\phi(\mathbf{k}, t)$ becomes

$$\begin{aligned} f_\phi(\mathbf{k}, t, \nu) &= \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \mathbb{E}\left[e^{-\theta_t S_t^\lambda} \theta_t^{S^k}\right] \\ &= \frac{K_p(\sqrt{ab})}{K_\nu(\phi)} c^\nu \left(\frac{b}{a}\right)^{\frac{p}{2}} \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \end{aligned} \tag{5.15}$$

where $a = \phi c + 2S_t^\lambda$, $b = \frac{\phi}{c}$ and $p = S^k + \nu$. Furthermore, we let ν be constant and fixed in order to avoid potential identification problems which may appear when performing estimation. In general, however, the derivative with respect to ϕ is really

hard to find since the constant c involves the Bessel function. On the other hand, it is worth noting that $var(\theta_t)$ is roughly unbounded when $\nu \in [-2, 0]$ and the skewness and kurtosis are decreasing with respect to ν, which can be easily verified by some statistical software on computer. So, we will discuss cases where $\nu = -\frac{1}{2}, -\frac{3}{2}, -\frac{3}{4}$, two of which have 'explicit' distributions in the sense that the constant c can be evaluated in closed form.

5.4.3.1 Generalized Inverse Gaussian with $\nu = -\frac{1}{2}$

In this case, the resulting distribution known as the Poisson Inverse Gaussian distribution is investigated by many authors (see,e.g Sichel (1974); STCHKL (1982),Atkinson and Yeh (1982), Stein and Juritz (1988) among others). When $\nu = -\frac{1}{2}$, $c = 1$ and the distribution function f becomes

$$f_\phi(\mathbf{k}, t) = \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \sqrt{\frac{2}{\pi}} \phi^{\frac{1}{2}} e^{\phi} K_p(\sqrt{\phi(\phi + 2S_t^\lambda)}) \left(\frac{\phi}{\phi + 2S_t^\lambda}\right)^{\frac{p}{2}} \tag{5.16}$$

For convenience, we reparametrize the above density by squaring the parameter ϕ such that

$$f_\phi(\mathbf{k}, t) = \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \sqrt{\frac{2}{\pi}} \phi e^{\phi^2} K_p(\phi\Delta)(\frac{\phi}{\Delta})^{\nu} \tag{5.17}$$

where $p = S^k - \frac{1}{2}$ and $\Delta = \sqrt{\phi^2 + 2S^\lambda}$. The derivatives of $f_\phi(\mathbf{k}, t)$ with respect to different parameters can be derived by making use of the derivative of $K_\nu(\omega)$ with respect to its argument such that

$$\frac{\partial K_\nu(\omega)}{\omega} = \frac{\nu}{\omega} K_\nu(\omega) - K_{\nu+1}(\omega), \tag{5.18}$$

then it leads to the following derivatives

$$\begin{aligned} \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \phi} &= \left(2\phi + \frac{1 + 2\nu}{\phi}\right) f_\phi(\mathbf{k}, t) - \left(\phi + \frac{\Delta^2}{\phi}\right) \frac{k_1 + 1}{\lambda_{1,t}} f_\phi(\mathbf{k} + \mathbf{1}_1, \phi), \\ \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \beta_i} &= \left(\frac{k_i}{\lambda_{i,t}} f_\phi(\mathbf{k}, t) - \frac{k_i + 1}{\lambda_{i,t}} f_\phi(\mathbf{k} + \mathbf{1}_i, \phi)\right) \lambda_{i,t} z_{i,t} \end{aligned} \tag{5.19}$$

where $\mathbf{1}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^{m \times 1}$ is vector with i-th element being one and 0 elsewhere.

5.4.3.2 Generalized Inverse Gaussian with $\nu = -\frac{3}{2}$

In this case, the constant $c = \frac{\phi}{1+\phi}$ and the variance $var(\theta_t) = \frac{1}{\phi}$ which is exactly the same as the variance of Inverse Gaussian case but the random effect θ_t will in general have larger skewness and kurtosis. The resulting distribution function is

$$f_\phi(\mathbf{k}, t) = \prod_{i=1}^{m} \frac{\lambda_{i,t}^{k_i}}{k_i!} \sqrt{\frac{2}{\pi}} (\phi+1)^{S^k - 1} e^{\phi} \omega^{-p} K_p(\omega). \tag{5.20}$$

where $p = S^k - \frac{3}{2}$ and $\omega = \sqrt{\phi^2 + 2(\phi+1)S_t^\lambda}$. The derivatives with respect to different parameters can be derived similar to that of Inverse Gaussian case

$$\begin{aligned} \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \phi} &= \left(\frac{\phi + S^k}{\phi + 1} \right) f_\phi(\mathbf{k}, t) - \left(\frac{k_1 + 1}{\lambda_{1,t}} \frac{\phi + S_t^\lambda}{\phi + 1} \right) f_\phi(\mathbf{k} + \mathbf{1}_1, t) \\ \frac{\partial f_\phi(\mathbf{k}, t)}{\partial \beta_i} &= \left(\frac{k_i}{\lambda_{i,t}} f_\phi(\mathbf{k}, t) - \frac{k_i + 1}{\lambda_{i,t}} f_\phi(\mathbf{k} + \mathbf{1}_i, \phi) \right) \lambda_{i,t} z_{i,t} \end{aligned} \tag{5.21}$$

The remaining case where $\nu = -\frac{3}{4}$ cannot be simplified since the $c = \frac{K_{1/4}(\phi)}{K_{3/4}(\phi)}$ cannot be written down in terms of basic functions. Hence numerical differentiation has to be applied when evaluating $\frac{\partial f_\phi(\mathbf{k},t)}{\partial \phi}$ and $\frac{\partial f_\phi(\mathbf{k},t)}{\partial \beta_i}$. Finally, Table 5.1 summarise the parametrization of all mixing densities and Table 5.2 shows the moments formula for each mixing density.

Table 5.1: Parametrization of mixing density based on GIG density 5.4

Mixing density	ψ	χ	ν	Range of parameter
Gamma	2ϕ	0	ϕ	$\phi > 0$
GIG with unit mean	$c\phi$	$\frac{\phi}{c}$	fixed constant in $\mathbb{R}$	$c = \frac{K_\nu(\phi+1)}{K_\nu(\phi)}, \phi > 0$
Inverse Gamma	0	2ϕ	$-\phi - 1$	$\phi > 1$

Table 5.2: Moments for the random effect θ_t. Ex.Kurtosis = Kurtosis - 3

Mixing density $g(\theta)$	Variance	Skenwness	Ex.Kurtosis
Gamma	$\frac{1}{\phi}$	$\frac{2}{\sqrt{\phi}}$	$\frac{6}{\phi}$
Inverse Gaussian	$\frac{1}{\phi^2}$	$\frac{3}{\phi}$	$\frac{15}{\phi^2}$
GIG $\nu = -\frac{3}{2}$	$\frac{1}{\phi}$	$\frac{\phi^{\frac{3}{2}} + \phi^{\frac{1}{2}}}{\phi^2}$	$\frac{3+12\phi+15\phi^2}{\phi^3}$
Inverse Gamma	$\frac{1}{\phi-1}$	$\frac{4\sqrt{\phi-1}}{\phi-2}$	$\frac{6(5\phi-1)}{(\phi-2)(\phi-3)}$

Although the formula for variances is slightly different due to its parametrization, they can be easily reparameterized and compared with each other. It turns out

that the Inverse Gamma has the largest skewness and kurtosis while the Gamma density has the smallest, which means the 'tailedness' of those density increases in a 'top-down' order according to the Table. Hence, one can choose different density to accommodate different tail structure encountered in real data.

5.5 Model fitting and Prediction

5.5.1 Maximum likelihood estimation for the MMPGIG-INAR(1) model

In this section, we derive the log likelihood function and score function of the MMPGIG-INAR(1) model defined above for the general case. Let the whole parameter set be $\Theta = \{p_i, \beta_i, \phi | i = 1, \ldots, m\}$ and then the log likelihood function $\ell(\Theta)$ for this discrete Markov chain is just the product of their conditional probability function such that $\ell(\Theta) = \prod_t \mathcal{P}_\Theta(\mathbf{X}_t|\mathbf{X}_{t-1})$, where the conditional probability is the convolution of m+1 distribution functions such that

$$\begin{aligned}
\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1}) &= \mathbb{E}\left[\prod_{i=1}^{2} \mathcal{P}(p_i \circ X_{i,t-1} + R_{i,t} = X_{i,t-1}|X_{i,t-1}, \theta_t)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{m} \sum_{k=0}^{s_i} f_{p_i}(k, X_{i,t-1}) f_{R_i}(X_{i,t} - k, t)\right], \quad s_i = \min\{X_{i,t-1}, X_{i,t}\} \\
&= \sum_{k_1=0}^{s_1} \cdots \sum_{k_m=0}^{s_m} f_\phi(\mathbf{k}, t) \prod_{i=1}^{m} f_{p_i}(X_{i,t} - k_i, X_{i,t-1}),
\end{aligned} \tag{5.22}$$

where the expectation is taken with respect to the random variable θ_t. The following proposition gives $\ell(\Theta)$ and its score functions.

Proposition 5.5. *Suppose there is a multivariate random sequence* $(\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_n)$ *generated from the MMPGIG-INAR(1) model, the log likelihood function* $\ell(\Theta)$ *and*

score functions are given by

$$\begin{aligned}
\ell(\Theta) &= \sum_{t=1}^{n} \log \mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1}) \\
&= \sum_{t=1}^{n} \log \sum_{k_1=0}^{s_1} \cdots \sum_{k_m=0}^{s_m} f_\phi(\mathbf{k}, t) \prod_{i=1}^{m} f_{p_i}(X_{i,t} - k_i, X_{i,t-1}) \qquad (5.23) \\
\frac{\partial \ell(\Theta)}{\partial \vartheta} &= \sum_{t=1}^{n} \frac{1}{\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})} \frac{\partial \mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}{\partial \vartheta}, \quad \vartheta \in \Theta
\end{aligned}$$

The derivatives inside the sum are given by

$$\begin{aligned}
\frac{\partial \mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}{\partial p_j} &= \sum_{k_1=0}^{s_1} \cdots \sum_{k_m=0}^{s_m} f_\phi(\mathbf{k}, t) \frac{\partial f_{p_j}(X_{ij,t} - k, X_{j,t-1})}{\partial p_j} \prod_{i \neq j} f_{p_j}(X_{i,t} - k_i, X_{i,t-1}) \\
\frac{\partial \mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}{\partial \vartheta_1} &= \sum_{k_1=0}^{s_1} \cdots \sum_{k_m=0}^{s_m} \frac{\partial f_\phi(\mathbf{k}, t)}{\vartheta_1} \prod_{i=1}^{m} f_{p_i}(X_{i,t} - k_i, X_{i,t-1}) \\
&\vartheta_1 \in \{\beta_1, \beta_2, \phi\}
\end{aligned} \qquad (5.24)$$

where the derivative $\frac{\partial f_{p_j}(\omega, X_{j,t-1})}{\partial p_j}$ *has the same form for all* $j = 1, ..., m$.

$$\frac{\partial f_{p_j}(\omega, X_{j,t})}{\partial p_j} = f_{p_j}(\omega, X_{j,t}) \frac{\omega - p_j X_{j,t}}{p_j(1 - p_j)}$$

The derivatives $\frac{\partial f_\phi(\mathbf{k}, t)}{\vartheta_1}$ are already discussed in Section 4 for different cases. Hence, the maximum likelihood estimators can be obtained through numerical algorithms, for example Newton-raphson, Quasi-Newton and so on. However, optimization will be computational intensive as m increases. One can solve this issue by adopting the composite likelihood method introduced in Pedeli and Karlis (2013a), where the high dimensional likelihood function was reduced to a sum of bivariate cases.

5.5.2 Integer-valued Prediction

Based on the estimates obtained by maximum likelihood and the random sequence $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, the h-steps ahead distribution of $\mathbf{X}_{n+h}$ conditional on $\mathbf{X}_n$ is given by

$$\mathbf{X}_{n+h} \stackrel{D}{=} \hat{\mathbf{P}}^h \circ \mathbf{X}_n + \sum_{k=1}^{h} \hat{\mathbf{P}}^{h-k} \circ \mathbf{R}_{n+k}, \qquad (5.25)$$

where $\hat{\mathbf{P}}$ is obtained from above estimation procedure. In the classical time series model, one would minimise $\text{MSE}(h) = \mathbb{E}[(\hat{\mathbf{X}}_{n+h} - \mathbf{X}_{n+h})^2 | \mathbf{X}_n]$ to obtain the optimal linear predictor such that $\hat{\mathbf{X}}_{n+h} = \mathbb{E}[\mathbf{X}_{n+h} | \mathbf{X}_n]$. However, this would inevitably introduce real value for $\hat{\mathbf{X}}_{n+h}$, which is not coherent to the integer-valued nature of MMPGIG-INAR(1) model. To solve this, one can instead use the median $\tilde{\mathbf{X}}_{n+h}$ of $\mathbf{X}_{n+h}$, the 50% quantile, as prediction value for the model, which is also discussed by Pavlopoulos and Karlis (2008) and Homburg et al. (2019). In the univariate case, the median is obtained by minimising the mean absolute error $\text{MAE}(h) = \mathbb{E}[|\tilde{X}_{n+h} - X_{n+h}||X_n]$. The idea here can be extended to the multivariate case so that the median $\tilde{\mathbf{X}}_{n+h}$ is called geometric median, which is calculated by minimising the expected Euclidean distance

$$MAE(h) = \mathbb{E}[||\tilde{\mathbf{X}}_{n+h} - \mathbf{X}_{n+h}||_2 | X_n] \tag{5.26}$$

On the other hand, the expectation can be evaluated numerically by simulating the random samples of $\mathbf{X}_{n+h}$.

5.6 Empirical analysis

The data used in this section come from the Local Government Property Insurance Fund (LGPIF) from the state of Wisconsin. This fund provides property insurance to different types of government units, which includes villages, cities, counties, towns and schools. The LGPIF contains three major groups of property insurance coverage, namely building and contents (BC), inland marine (IM) and motor vehicles (PN, PO, CN, CO). For exploratory purposes, we focus on modelling jointly the claim frequency of IM, denoted as $X_{1,t}$, and comprehensive new vehicles collision (CN), denoted as $X_{2,t}$. The insurance data cover the period over 2006 - 2011 with 1234 policyholder records in total. Only $n_1 = 1048$ of them have complete data over the period 2006-2010 which will be used as the training data set. The last year 2011 with $n_2 = 1025$ policyholders out of 1048 in the data set will be the test data set. Denote the IM type and CN type claim frequency for a particular policyholder as $X_{1,t}^{(j)}, X_{2,t}^{(j)}$ respectively, where j is the identifier for each policyholder. Then the relationship between $X_{i,t}$ and $X_{i,t}^{(j)}$ is simply $X_{i,t} = \sum_{j=1}^{n_1} X_{i,t}^{(j)}$ with $i = 1, 2$ while t would take the values from 1 to 5 corresponding to the year 2006 to 2010.

In what follows, basic statistical analysis is shown in Table 5.3 and figures 5.1 and 5.2. The proportion of zeros for the two types of claims is higher than 90% during the period 2006-2010. Also, both types of claims exhibit overdispersion, since their

variances exceeds their means during this period. Furthermore, the overdispersion for $X_{2,t}$ is even stronger than that of $X_{1,t}$, which indicates the need to employ an overdispersed distribution for this data. Additionally, the correlation tests for $X_{1,t}$ and $X_{2,t}$ show a positive correlation between the two claim types. At this point it is worth noting that modelling positively correlated claims has been explored by many articles. See for example, Bermúdez and Karlis (2011), Bermúdez and Karlis (2012), Shi and Valdez (2014a), Shi and Valdez (2014b), Abdallah et al. (2016), Bermúdez and Karlis (2017), Bermúdez et al. (2018), Bermúdez et al. (2018), Pechon et al. (2018), Pechon et al. (2019), Bolancé and Vernic (2019), Denuit et al. (2019), Fung et al. (2019), Bolancé et al. (2020), Pechon et al. (2021), Jeong and Dey (2021), Gómez-Déniz and Calderín-Ojeda (2021), Tzougas and di Cerchiara (2023), Tzougas and di Cerchiara (2021) and Bermúdez and Karlis (2021). Finally, the proportion of zeros and kurtosis show that the marginal distributions of $X_{1,t}, X_{2,t}$ are positively skewed and exhibit a fat-tailed structure which indicates the appropriateness of adopting a positive skewed and fat-tailed distribution (GIG distribution).

Table 5.3: Summary statistics of two types of claims over years. The correlations test is a one-sided test where the alternative hypothesis is "The sample correlation is greater than 0"

	2006	2007	2008	2009	2010
Proportion of zeros $X_{1,t}$	0.9685	0.9542	0.9552	0.9504	0.9590
Proportion of zeros $X_{2,t}$	0.9342	0.9332	0.9399	0.9370	0.9323
Kurtosis of $X_{1,t}$	85.75009	86.64794	41.84915	43.01832	126.68793
Kurtosis of $X_{2,t}$	53.91835	61.77408	111.28108	184.13950	133.92283
P-value of correlation test	0.0000	0.0000	0.0000	0.0000	0.0000

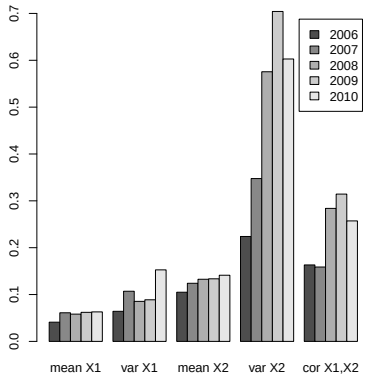


Figure 5.1: summary statistics (mean, variance and correlation) for each type of claims across all the policyholders over years.

The description and some summary statistics for all the explanatory variables (covariates $z_{1,t}, z_{2,t}$) that are relevant to $X_{1,t}, X_{2,t}$ are shown in Table 5.4. Variables

1- 5 including 'TypeVillage' are categorical variables to indicate the entity types of a policyholder. Due to the strongly heavy-tailed structure appearing in variables 6 and 9 which can drastically distort the model fitting, those variables are transformed by means of the 'rank' function in R software and then standardized, which can mitigate the effect of outliers. Variables 6-8 are relevant to IM claim $X_{1,t}$ while variables 9,10 provide information for CN claims $X_{2,t}$. The covariate $z_{1,t}$ includes variables 1-8 and $z_{2,t}$ contains variables 1-5 and variables 9,10. These covariates act as the regression part for $\lambda_{i,t}$ mentioned in section 2, which may help explained part of the heterogeneity between $X_{1,t}$ and $X_{2,t}$.

Table 5.4: summary statistics for the explanatory variables

Variable index	Variable name	Type	Description	Proportion/Mean
1	TypeCity	Categorical	Indicator for city entity	0.1400
2	TypeCounty	Categorical	Indicator for county entity	0.0578
3	TypeMisc	Categorical	Indicator for miscellaneous entity	0.1104
4	TypeSchool	Categorical	Indicator for school entity	0.2817
5	TypeTown	Categorical	Indicator for town entity	0.1728
-	TypeVillage	Categorical	Indicator for village entity (reference category)	0.2373
6	CoverageIM	Continuous	Coverage amount of IM(transformed)	0
7	InDeductIM	Continuous	Log deductible amount for inland marine	5.3400
8	NoClaimCreditIM	Binary	Indicator for no IM claims in prior year	0.4210
9	CoverageCN	Countinuous	Coverage amount of CN (transformed)	0
10	NoClaimCreditCN	Binary	Indicator for no CN claims in prior year	0.0897

The MMPGIG-INAR(1) with $m = 2$, is applied to model the joint behaviour of $X_{1,t}^{(j)}, X_{2,t}^{(j)}$ across all the policyholders. Note that when Gamma mixing density is used in MPGIG INAR(1), the resulting model will be the "BINAR(1) Process with BVNB Innovations" in Pedeli and Karlis (2011), which we will used as comparison benchmark for other choices of mixing density. The the likelihood function would simply become

$$\ell(\Theta) = \sum_{j=1}^{n_1} \ell_j(\Theta) = \sum_{j=1}^{n_1} \sum_{t=1}^{4} \log \Pr(X_{1,t+1}^{(j)}, X_{2,t+1}^{(j)} | X_{1,t}^{(j)}, X_{2,t}^{(j)}), \tag{5.27}$$

where $\ell_j(\Theta)$ is the likelihood function for policyholder j. Note that all the policyholders with the same type of claim $X_{i,.}$ will share the same set of parameters p_i, β_i and ϕ will be same for both claim types. In addition, it is necessary to show the appropriateness of introducing correlation and time-series component (binomial thinning) in MPGIG INAR(1). Then we also fit the data to following models.

1. The joint distribution of $X_{1,t}^{(j)}$ and $X_{2,t}^{(j)}$ are assumed to be bivariate mixed Poisson distribution (BMP) with probability mass function $f_\phi(\mathbf{k}, t)$ which we already discussed in section 4.

2. The joint distribution of $X_{1,t}^{(j)}$ and $X_{2,t}^{(j)}$ are characterized by two independent INAR(1) models (TINAR)

$$X_{1,t}^{(j)} = p_1 \circ X_{1,t-1}^{(j)} + R_{1,t}$$
$$X_{2,t}^{(j)} = p_2 \circ X_{2,t-1}^{(j)} + R_{2,t},$$

where $R_{i,t} \sim Pois(\lambda_{i,t}\theta_{i,t}), i = 1, 2$ and random effect $\theta_{i,t}$ is independent of i.

Similarly, the likelihood functions for these models will have the same form as equation 5.27 but different joint distribution $\Pr(X_{1,t+1}^{(j)}, X_{2,t+1}^{(j)} | X_{1,t}^{(j)}, X_{2,t}^{(j)})$. For comparison purposes, we fit the bivariate Poisson mixture regression model with the training data starting from 2007 because BMP model does not need to consider lag responses.

All the estimations is implemented in R software by the 'optim' function with method 'BFGS' (quasi-Newton method). The gradient functions with respect to all the parameters are derived in section 4 and section 5 and they can be input as gradient argument in 'optim' function, which will significantly decrease the amount of computational time compared to numerical gradient function in default setting.

Table 5.5: The AIC and BIC when fitting as two independent INAR model with different combination. row = Gamma, column = Inverse Gaussian means that gamma mixing density for $X_{1,t}$ and Inverse Gaussian mixing density for $X_{2,t}$. For each entry, the number on the left is AIC and the other one is BIC.

Mixing density	Gamma	Inverse Gaussian	GIG $\nu = -\frac{3}{4}$	GIG $\nu = -\frac{3}{2}$	Inverse Gamma
	AIC | BIC	AIC | BIC	AIC | BIC	AIC | BIC	AIC | BIC
Gamma	2999.957 | 3133.117	2999.433 | 3132.592	2999.590 | 3132.749	3000.48 | 3133.64	3002.352 | 3135.512
Inverse Gaussian	2998.876 | 3132.036	2998.351 | 3131.511	2998.508 | 3131.668	2999.399 | 3132.558	3001.271 | 3134.430
GIG $\nu = -\frac{3}{4}$	2998.661 | 3131.820	2998.136 | 3131.296	2998.293 | 3131.453	2999.184 | 3132.343	3001.056 | 3134.215
GIG $\nu = -\frac{3}{2}$	3004.415 | 3137.574	3003.890 | 3137.050	3004.047 | 3137.207	3004.938 | 3138.097	3006.810 | 3139.969
Inverse Gamma	2998.778 | 3131.938	2998.254 | 3131.413	2998.410 | 3131.570	2999.301 | 3132.461	3001.173 | 3134.333

Table 5.6: The AIC and BIC when fitting bivariate sequence as bivarate mixed Poisson regression model and BINAR model. For each entry, the number on the left is AIC and the other one is BIC.

Mixing density	Gamma	Inverse Gaussian	GIG $\nu = -\frac{3}{4}$	GIG $\nu = -\frac{3}{2}$	Inverse Gamma
	AIC | BIC	AIC | BIC	AIC | BIC	AIC | BIC	AIC | BIC
BMP	3073.149 | 3187.285	3061.892 | 3176.028	3060.961 | 3175.098	3079.179 | 3193.315	3059.374 | 3173.510
BINAR(1)	2996.291 | 3123.109	2992.953 | 3119.771	2992.854 | 3119.672	3008.458 | 3135.277	2995.348 | 3122.167

Model fitting results are shown in Tables 5.5 and 5.6. All the results show a great improvement by adopting a time series model compared to BMP results in Table 5.6. When focusing on the results of BINAR in Table 5.6, except the case where the mixing density is GIG $\nu = -\frac{3}{2}$, there is an significant improvement by introducing

the fat-tailed distribution as mixing density in $\mathbf{R}_t$ compared to Gamma case. On the other hand, the improvement from the optimal TINAR to the optimal BINAR (cells are in grey color) is obvious, which is indicated by lower AIC and BIC of BINAR with GIG $\nu = -\frac{3}{4}$ compared to TINAR with GIG $\nu = -\frac{3}{4}$ and Inverse Gaussian. It implies that there is significant correlation between two claim sequences. Maximum likelihood estimates for three cases are given in Table 5.7 as well as their standard deviations. The standard derivations are estimated by inverting the numerical Hessian matrix. From Table 5.7 we see that the estimates for p_i, β_i are very close to each other while the estimated ϕ is significantly different among three mixing densities, which is expected because ϕ influences the tail and correlation structure of the bivariate sequence $X_{1,t}, X_{2,t}$. Furthermore, we see that the explanatory variables have a similar effect (positive and/or negative) and are almost identical for both response variables in the case of of all three models. Finally, the variables which are statistically significant at a 5 % threshold for $X_{1,t}$ are TypeCounty, TypeMisc, TypeVillage, NoClaimCreditIM, and those which are statistically significant at a 5 % threshold for $X_{2,t}$ are TypeCity, TypeCounty, TypeVillage, CoverageIM, and CoverageCN.

The Figure 5.2 below presents prediction for both types of claims at $t = 2011$ with $n_2 = 1025$ policyholders based on geometric median equation 5.26. It seems that the prediction for number of policyholders who make no claims are reasonably good while the prediction for $X_{1,t}$ are generally underestimated at tail and the prediction for $X_{2,t}$ are overestimated at the tail. On the other hand, Table 5.8 shows the prediction sum of squared error (PSSE) and frequency of some basic combination of observations, namely $(0,0), (1,0), (0,1), (1,1)$ for the best fitted models within three classes, bivariate mixed Poisson regression, Two independent INAR(1) and bivariate INAR(1). It is again clear that the introduction of autoregressive part makes sense as it greatly reduce the prediction error. Although the best TINAR model has the closet frequency of $(0,0)$, the best BINAR model has the lowest overall prediction error.

Table 5.7: Maximum likelihood estimation for MMPGIG-INAR(1) of insurance's claim frequency data when $m = 2$. For each entry, the upper one is the estimate and the estimated standard deviations are indicated in square brackets

Estimate \ Mixing density	Gamma		Inverse Gaussian		GIG $\nu = -\frac{3}{4}$	
	$X_{1,t}$	$X_{2,t}$	$X_{1,t}$	$X_{2,t}$	$X_{1,t}$	$X_{2,t}$
p_i	0.1238	0.2904	0.1200	0.2768	0.1194	0.2750
	(0.0373)	(0.0378)	(0.0376)	(0.0388)	(0.0376)	(0.0388)
ϕ	0.9495		0.8885		0.7944	
	(0.1662)		(0.0931)		(0.0136)	
Intercept	-3.7980	-5.9744	-3.8228	-5.9967	-3.8287	-6.0029
	(0.5158)	(0.4428)	(0.5206)	(0.4503)	(0.5213)	(0.4489)
TypeCity	-0.2242	0.6673	-0.2316	0.6510	-0.2337	0.6480
	(0.2555)	(0.2823)	(0.2577)	(0..2861)	(0.2586)	(0.2866)
TypeCounty	0.5682	1.3290	0.5784	1.3019	0.5814	1.2953
	(0.2811)	(0.2643)	(0.2836)	(0.2674)	(0.2794)	(0.2673)
TypeMisc	-2.0110	-0.1141	-2.0213	-0.1313	-2.0226	-0.1342
	(1.0210)	(0.6567)	(1.0223)	(0.6592)	(1.0244)	(0.6577)
TypeSchool	-0.0387	0.1559	-0.0638	0.1323	-0.0692	0.1279
	(0.3587)	(0.2811)	(0.3570)	(0.2837)	(0.3534)	(0.2841)
TypeTown	-0.3565	-0.8941	-0.3661	-0.9155	-0.3680	-0.9195
	(0.3037)	(0.4794)	(0.3048)	(0.4816)	(0.3054)	(0.4820)
CoverageIM	1.4543		1.4309		1.4259	
	(0.2126)		(0.2115)		(0.2080)	
InDeductIM	0.0170		0.0243		0.0259	
	(0.0788)		(0.0795)		(0.0792)	
NoClaimCreaditIM	-0.4569		-0.4501		-0.4482	
	(0.1570)		(0.1579)		(0.1579)	
CoverageCN		2.4227		2.4596		2.4675
		(0.2210)		(0.2260)		(0.2249)
NoClaimCreaditCN		-0.3047		-0.3231		-0.3261
		(0.1811)		(0.1814)		(0.1801)

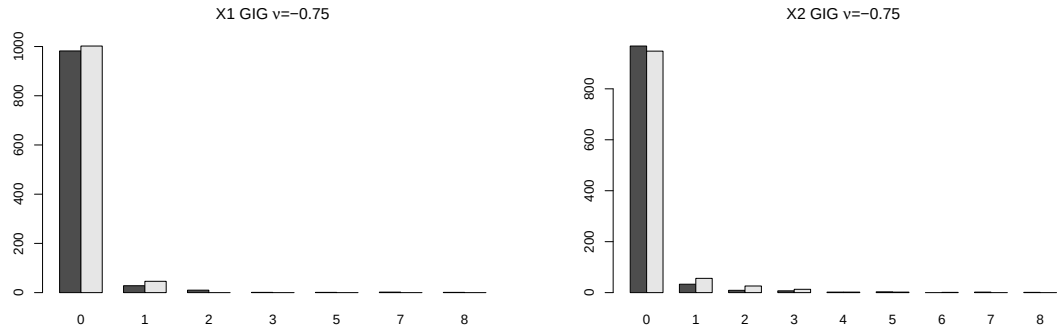


Figure 5.2: Observed (dark) and Predicted(grey) frequency of the test data set based on estimated BINAR with GIG $\nu = -\frac{3}{4}$ as mixing density

Table 5.8: Summary of prediction on test data

Model	PSSE			Frequency			
	X_1	X_2	$X_1 + X_2$	(0,0)	(1,0)	(0,1)	(1,1)
BMP Inverse gamma	264	221	485	970	0	24	0
TINAR GIG $\nu = -0.75$ & Inverse gamma	212	164	376	964	4	33	0
BINAR GIG $\nu = -0.75$	212	156	368	966	3	32	0
Observed				940	20	26	6

5.7 Concluding remarks

In this paper we proposed the MMPGIG-INAR(1) regression model for modelling multiple time series of different types of count response variables. The proposed model, which is an extension of BINAR(1) regression model that was introduced by Pedeli and Karlis (2011), can accommodate positive correlation and multivariate overdispersion in a flexible manner. In particular, the Generalized Inverse Gaussian class includes many distributions as its special and limiting cases that can be used for modelling the innovations $\mathbf{R}_t$. Thus, the proposed modelling framework can efficiently capture the stylized characteristics of alternative complex data sets. Furthermore, due to the simple form of its density function, statistical inference for the MMPGIG-INAR(1) model is straightforward via the ML method, whereas other models that have been proposed in the literature, such as copula-based models, may result in numerical instability during the ML estimation procedure. For demonstration purposes different members of the proposed famly of models were fitted to LGPIF data from the state of Wisconsin. Finally, it is worth mentioning that a possible line of further research could be to also consider cross correlation, meaning that the non-diagonal elements of $\mathbf{P}$ can take positive values.

CHAPTER 6

Paper C. EM Estimation for Bivariate Mixed Poisson INAR(1) Claim Count Regression Models with Correlated Random Effects

Abstract

This article considers bivariate mixed Poisson INAR(1) regression models with correlated random effects for modelling correlations of different signs and magnitude among time series of different types of claim counts. This is the first time that the proposed family of INAR(1) models is used in a statistical or actuarial context. For expository purposes, the bivariate mixed Poisson INAR(1) claim count regression models with correlated Lognormal and Gamma random effects paired via a Gaussian copula are presented as competitive alternatives to the classical bivariate Negative Binomial INAR(1) claim count regression model which only allows for positive dependence between the time series of claim count responses. Our main achievement is that we develop novel alternative Expectation-Maximization type algorithms for maximum likelihood estimation of the parameters of the models which are demonstrated to perform satisfactory when the models are fitted to Local Government Property Insurance Fund data from the state of Wisconsin.

6.1 Introduction

Over the past decade, there has been a growing literature on bivariate (and/or multivariate) claim count regression models which can efficiently capture the dependence between claims from the same policy and/or different coverages bundled into a single policy. The interested reader is referred to Abdallah et al. (2016), Bermúdez and Karlis (2011), Bermúdez and Karlis (2012), Bermúdez and Karlis (2021), Bermúdez et al. (2018), Bolancé et al. (2020), Bolancé and Vernic (2019), Denuit et al. (2019), Fung et al. (2019), Gómez-Déniz and Calderín-Ojeda (2021), Jeong and Dey (2021), Pechon et al. (2019), Pechon et al. (2021), Pechon et al. (2018), Shi and Valdez (2014a), and Shi and Valdez (2014b) among many others.

Pechon et al. (2018) proposed the use of bivariate mixed Poisson count regression models, with correlated random effects for capturing the interactions between the different coverages purchased by members of the same household. In particular, Pechon et al. (2018) considered the bivariate Poisson-Gamma (BPGGA) regression model with Gaussian copula and the bivariate Poisson-Lognormal (BPLN) regression model. In the former model the random effects are distributed according to two Gamma distributions with unit means and the dependence between the random effects is introduced by means of a Gaussian bivariate copula whereas in the latter model these random effects are distributed according to the bivariate Lognormal mixing distribution. Bermúdez et al. (2018), following the setup of Pedeli and Karlis (Pedeli and Karlis (2011) and Pedeli and Karlis (2013a)), were the first to derive a bivariate Poisson integer-valued autoregressive process of order 1 (BINAR(1)) claim count regression model which can account both for cross-sectional and temporal dependence between multiple claim types. The model they developed was employed for addressing the ratemaking problem of pricing an insurance contract in the case of positively correlated claims from different types of coverage in non-life insurance. Finally, Bermúdez and Karlis (2021) built on the previous paper by using a multivariate INAR(1) (MINAR(1)) regression model based on the Sarmanov family of distributions. The MINAR(1) regression models based on the Sarmanov family of distributions are also restricted to a positive correlation structure between the claim count response variables. However, it enjoys some advantages compared to a different approach which can allow for both positive and negative correlations by using copulas for the specification of the joint distribution of the innovations. See, for instance, Cameron et al. (2004), Karlis and Pedeli (2013), Lee (1999), Nikoloulopoulos (2013), Nikoloulopoulos (2016) and Nikoloulopoulos and Karlis (2010) among others. Firstly, it avoids identifiability issues which may arise when a continuous copula distribution is paired with discrete marginals, see Genest and Nešlehová (2007). As

is well known, the lack of identifiability means that it cannot be guaranteed that model fitting is unique and this may lead to problems in statistical inference, for example, one might receive no meaningful values for the standard errors of the parameters. Secondly, the computational intensity for discrete copula-based models increases as the dimension of the model increases and hence, as is also mentioned by the authors, their approach, which relies on the use of the Sarmanov family, provides models that are less computationally intensive to estimate and can still have a reasonable range for positive dependence structure between the claim count responses.

In this study, we introduce a family of bivariate mixed Poisson INAR(1) claim count regression models with correlated random effects for modelling the dependence structure between times series of different types of claim counts from the same and/or different types of coverage. The bivariate mixed Poisson INAR(1) regression models with correlated random effects are a broad class of models which can accommodate overdispersion, which is a direct consequence of unobserved heterogeneity due to systematic effects in the data, and correlations of different signs and magnitude. For demonstration purposes, we consider the bivariate mixed Poisson INAR(1) claim count regression models which are derived by using the bivariate Lognormal and Gaussian copula paired with gamma marginals as mixing densities, which we refer to as BINAR(1)-LN and BINAR(1)-GGA claim count regression models respectively. Both models can be regarded as extensions of the classical bivariate Negative Binomial INAR(1) claim count regression model with a shared gamma random effect, which we refer to as BINAR(1)-GA claim count regression model, in the sense that they provide more flexibility for modelling overdispersed bivariate time series of count data compared to the BINAR(1)-GA model which is derived by pre-imposing the restrictive positive correlation assumption between time series of different claim types of claim counts, since in some cases negative correlations may be of interest as well. Furthermore, unlike previous copula-based count regression models for which identifiability issues can arise when a continuous copula distribution is paired with discrete marginals, in the proposed family of models identifiability of the bivariate distribution of the innovations is guaranteed by imposing a unit mean constraint for the Gamma continuous mixing densities which are paired with a Gaussian copula.

The main contributions we make are as follows:

- Firstly, before we introduce the time series components, we present a unified framework for statistical inference via the Expectation-Maximization (EM) algorithm for the BPGA, BPLN and BPGGA regression models[1].

[1]Note that EM estimation for the parameters of the BPGA regression model with a shared

- Secondly, we develop novel EM type algorithms for maximum likelihood (ML) estimation of the BINAR(1)-GA, BINAR(1)-LN and BINAR(1)-GGA regression models, which has not been explored in the literature so far. The main reason for this is because the joint distribution of the innovations cannot be written in closed form in either model and hence its maximization is not possible via standard numerical optimization as is done in Bermúdez et al. (2018), Bermúdez and Karlis (2021), Karlis and Pedeli (2013), Pedeli and Karlis (2011) and Pedeli and Karlis (2013a)

The rest of the paper is organized as follows. Section 2 presents the model specifications for the bivariate mixed Poisson regression models we consider and describes their ML estimation via the EM algorithm. Section 3 presents the derivation of their INAR(1) extensions that we first proposed herein and outlines the EM type algorithms we developed for statistical inference. Section 4 presents our empirical analysis which is based on the LGPIF dataset. Concluding remarks are given in Section 5. The interpretation of abbreviations used in the paper and some other technical details are provided in appendix 6.A.

6.2 The bivariate mixed Poisson regression model

6.2.1 Model specifications

The bivariate Poisson mixture is constructed by two independent Poisson random variables conditional on a random effect vector (or scalar) $\boldsymbol{\theta} = (\theta_1, \theta_2)$ such that $N^{(i)} \sim Pois(\lambda_{i,t}\theta_i),\ i = 1, 2$. The bivariate mixed Poisson regression is then constructed by further allowing the rate λ_i to be modelled as functions of explanatory variables $\mathbf{z}_{i,t}$ such that $\lambda_{i,t} = \exp\{\mathbf{z}_{i,t}^T \boldsymbol{\beta}_i\}$. Denote the mixing density function of the random effect as $f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ parametrized by $\boldsymbol{\phi}$. To avoid the identifiability issue, we have to restrict the expectation $\mathbb{E}[\theta_i]$ to be a fixed constant. One usually lets $\mathbb{E}[\theta_i] = 1$ so that $\boldsymbol{\lambda}_t := (\lambda_{1,t}, \lambda_{2,t})$ will fully explain the frequency of a event and $\boldsymbol{\phi}$ will explain the variation and correlation of the whole bivariate sequence. In the following, we will discuss three different mixing densities, univariate gamma (shared random effect), bivariate Lognormal and Gaussian copula paired with Gamma marginals.

random effect and the BPLN regression model has been discussed in Gurmu and Elder (2000) and Silva et al. (2019) respectively. However, this is the first time that the EM algorithm is used for estimating the parameters of the BPGA regression model with Gaussian copula.

(a) **Univariate Gamma density**

In this case, the bivariate mixed Poisson regression model shares the same random effect $N_t^{(i)} \sim Pois(\lambda_{i,t}\theta)$ $i = 1, 2$. Denote the mixing density function as $f_\phi(\theta) = f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and it has following expression

$$f_\phi(\theta) = \frac{\phi^\phi}{\Gamma(\phi)}\theta^{\phi-1}e^{-\phi\theta}, \tag{6.1}$$

which has unit mean and variance $\frac{1}{\phi}$. Then the unconditional probability mass function $f_{PG}(\mathbf{k}, t)$ of $\mathbf{N}_t := (N_t^{(1)}, N_t^{(2)})$ can be written down in a closed form

$$\begin{aligned} f_{PG}(\mathbf{k}, t) &= \frac{\lambda_{1,t}}{k_1!}\frac{\lambda_{2,t}}{k_2!}\int_0^\infty e^{-(\lambda_{1,t}+\lambda_{2,t})\theta}\theta^{k_1+k_2}f_\phi(\theta)d\theta \\ &= \frac{\Gamma(\phi+k_1+k_2)}{\Gamma(\phi)\Gamma(k_1+1)\Gamma(k_2+1)}\frac{\phi^\phi\lambda_{1,t}^{k_1}\lambda_{2,t}^{k_2}}{(\phi+\lambda_{1,t}+\lambda_{2,t})^{\phi+k_1+k_2}}. \end{aligned} \tag{6.2}$$

(b) **Bivariate Lognormal density**

Suppose now the random vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$ follows bivariate normal distribution, with mean vector $(-\frac{\phi_1^2}{2}, -\frac{\phi_2^2}{2})$ and covariance matrix Σ

$$\Sigma = \begin{bmatrix} \phi_1^2 & \rho\phi_1\phi_2 \\ \rho\phi_1\phi_2 & \phi_2^2 \end{bmatrix} \tag{6.3}$$

Then the random effect vector $\boldsymbol{\theta} = e^{\boldsymbol{\epsilon}} = (e^{\epsilon_1}, e^{\epsilon_2})$ has Lognormal distribution with unit mean. Denote the density function of $\boldsymbol{\epsilon}$ as f_Σ^N and f_Σ^{LN} for Lognormal density. Then they have the following expressions

$$\begin{aligned} f_\Sigma^N(\boldsymbol{\epsilon}) =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\epsilon_1+0.5\sigma_1^2}{\sigma_1}\right)^2 - 2\rho\left(\frac{\epsilon_1+0.5\sigma_1^2}{\sigma_1}\right)\left(\frac{\epsilon_2+0.5\sigma_2^2}{\sigma_2}\right) + \left(\frac{\epsilon_2+0.5\sigma_2^2}{\sigma_2}\right)^2\right]\right\} \\ f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) =& \frac{1}{\theta_1\theta_2}f_\Sigma^N(\log\boldsymbol{\theta}) = f_\Sigma^{LN}(\boldsymbol{\theta}). \end{aligned}$$

The unconditional distribution $f_{PLN}(\mathbf{k}, t)$ of $\mathbf{N}_t$ is expressed as a double integral

$$\begin{aligned} f_{PLN}(\mathbf{k}, t) &= \int_0^\infty\int_0^\infty \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!}e^{-\lambda_{1,t}\theta_1}e^{-\lambda_{2,t}\theta_2}\theta_1^{k_1}\theta_2^{k_2}f_\Sigma^{LN}(\boldsymbol{\theta})d\theta_1 d\theta_2 \\ &= \int_R\int_R \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!}\exp\{-\lambda_{1,t}e^{\epsilon_1} - \lambda_{2,t}e^{\epsilon_2} + k_1\epsilon_1 + k_2\epsilon_2\}f_\Sigma^N(\boldsymbol{\epsilon})d\epsilon_1 d\epsilon_2. \end{aligned} \tag{6.4}$$

All the double integrals with respect to Lognormal density f_Σ^{LN} can be transformed into double integrals with respect to normal density f_Σ^N so that they can be evaluated by Gauss-Hermite quadrature. See details in the appendix 6.B .

(c) **Gaussian copula paired with Gamma marginals**

Suppose now the random vector $\boldsymbol{\theta}$ is distributed as a meta Gaussian copula such that its marginals are two independent Gamma random variables with parameter (ϕ_1, ϕ_2) respectively. Define uniform random vector $\mathbf{u} = (F_{\phi_1}(\theta_1), F_{\phi_2}(\theta_2))$. The distribution function $F_{GC}(\boldsymbol{\theta})$ and density function $f_{GC}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} F_{GC}(\boldsymbol{\theta}) &= C_\rho(\mathbf{u}) = F_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \\ f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) &= f_{GC}(\boldsymbol{\theta}) = \frac{f_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))}{f_{sn}(\Phi^{-1}(u_1)) f_{sn}(\Phi^{-1}(u_1))} f_{\phi_1}(\theta_1) f_{\phi_2}(\theta_2) \\ &:= c_\rho(\mathbf{u}) f_{\phi_1}(\theta_1) f_{\phi_2}(\theta_2), \end{aligned} \tag{6.5}$$

where $f_\rho(.,.), F_\rho(.,.)$ are the density function and cumulative distribution of bivariate normal random variable with the following expression

$$f_\rho(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2}\right\}. \tag{6.6}$$

The $\Phi(x)$ is the cdf of standard normal random variable with $\Phi^{-1}(x)$ as its quantile function and $f_{sn}(x)$ is the density function of the standard normal random variable. Finally, $f_{\phi_i}(x)$ and $F_{\phi_i}(x)$ are the pdf and cdf of Gamma density function defined in 6.1 for $i = 1, 2$. Then a bivariate Poisson Gamma random vector is constructed as $N_t^{(i)} \sim Pois(\lambda_{i,t}\theta_i), i = 1, 2$ with probability mass function $f_{PGC}(\mathbf{k}, t)$ such that

$$\begin{aligned} f_{PGC}(\mathbf{k}, t) &= \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!} \int_0^\infty \int_0^\infty \exp\{-\lambda_{1,t}\theta_1 - \lambda_{2,t}\theta_2\} \theta_1^{k_1}\theta_2^{k_2} f_{GC}(\theta_1, \theta_2) d\theta_1 d\theta_2 \\ &= \frac{\lambda_{1,t}^{k_1}}{k_1!}\frac{\lambda_{2,t}^{k_2}}{k_2!} \int_0^1 \int_0^1 e^{-\lambda_{1,t}F_{\phi_1}^{-1}(u_1) - \lambda_{2,t}F_{\phi_2}^{-1}(u_2)} F_{\phi_1}^{-1}(u_1)^{k_1} F_{\phi_2}^{-1}(u_2)^{k_2} c_\rho(u_1, u_2) du_1 du_2 \end{aligned}$$

Then the double integral can be evaluated by Gauss-Legendre quadrature. See details in appendix 6.C.

6.2.2 The EM algorithm

For statistical inference of above model, the classical maximum likelihood estimation is not straightforward to apply because the log likelihood function

$$\ell(\Theta) = \sum_{t=1}^{n} \log \left(\frac{\lambda_{1,t}^{k_{1,t}} \lambda_{2,t}^{k_{2,t}}}{k_{1,t}! k_{2,t}!} \int_0^\infty \int_0^\infty \theta_1^{k_{1,t}} \theta_2^{k_{2,t}} e^{-\lambda_{1,t}\theta_1 - \lambda_{2,t}\theta_2} f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) d\theta_1 d\theta_2 \right) \tag{6.7}$$

is not computational tractable and its maximum likelihood estimators are not straightforward to achieve. Alternatively, we can apply the EM algorithm to estimate the parameters $\Theta = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi}\}$. For given random samples $(\mathbf{k}_1, ...\mathbf{k}_n)$, suppose now we observe the random effect $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, then the complete likelihood function $\ell_c(\Theta)$ is given by

$$\ell_c(\Theta) = \sum_{t=1}^{n} \left[\left(\sum_{i=1}^{2} k_{i,t} log(\lambda_{i,t}\theta_{i,t}) - \lambda_{i,t}\theta_{i,t} - \log(k_{i,t}!) \right) + \log f_{\boldsymbol{\phi}}(\boldsymbol{\theta}_t) \right] \tag{6.8}$$

Compared to $\ell(\Theta)$, the complete log likelihood function $\ell_c(\Theta)$ are simplified in the sense that there is no integration and mixture likelihood are decomposed into Poisson likelihood and the likelihood for mixing density.

However, to evaluate $\ell_c(\Theta)$ we need to find out the conditional (posterior) distribution of $\boldsymbol{\theta}$ given the random samples. Then we define $\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) = e^{-\lambda_{1,t}\theta_1 - \lambda_{2,t}\theta_2} \theta_1^{k_{1,t}} \theta_2^{k_{2,t}}$ and posterior density

$$\begin{aligned} \pi(\boldsymbol{\theta}|\Theta^{(j)}, \mathbf{k}_t) &= \frac{f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \prod_{i=1}^{2} f_{Po}(k_{i,t}|\lambda_{i,t}\theta_i)}{\int_0^\infty \int_0^\infty f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \prod_{i=1}^{2} f_{Po}^{(j)}(k_{i,t}|\lambda_{i,t}\theta_i) d\theta_1 d\theta_2} \\ &= \frac{\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_{\boldsymbol{\phi}}(\boldsymbol{\theta})}{\int_0^\infty \int_0^\infty \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) d\theta_1 d\theta_2}. \end{aligned} \tag{6.9}$$

Then posterior expectation for any real value function $h(\boldsymbol{\theta})$ is given by

$$\begin{aligned} \mathbb{E}[h(\boldsymbol{\theta})|\Theta^{(j)}, \mathbf{k}_t] &= \int_0^\infty \int_0^\infty h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\Theta^{(j)}, \mathbf{k}_t) d\theta_1 d\theta_2 \\ &=: \mathbb{E}_{\boldsymbol{\theta},t}^{(j)}[h(\boldsymbol{\theta})], \end{aligned} \tag{6.10}$$

where $f_{Po}(k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$ is the probability mass function of a Poisson random variable with rate λ and the condition $\Theta^{(j)}$ means that the posterior density function is evaluated with the parameters estimated at j-th iteration. The subscript $\boldsymbol{\theta}$ of $\mathbb{E}_{\boldsymbol{\theta},t}^{(j)}$ means that the expectation is taken with respect to the $\boldsymbol{\theta}$ for t-th observation.

- **E-step:** Evaluating the Q function $Q(\Theta;\Theta^{(j)})$ given the the parameters estimated at j-th iteration

$$\begin{aligned} Q(\Theta;\Theta^{(j)}) \propto & \sum_{t=1}^{n}\sum_{i=1}^{2} k_{i,t}\log(\lambda_{i,t}) - \lambda_{i,t}\mathbb{E}[\theta_i|\Theta^{(j)},\mathbf{k}_t] + \sum_{t=1}^{n}\mathbb{E}[\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta})|\Theta^{(j)},\mathbf{k}_t] \\ = & \sum_{t=1}^{n}\sum_{i=1}^{2} k_{i,t}\log(\lambda_{i,t}) - \lambda_{i,t}\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\theta_i] + \sum_{t=1}^{n}\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta})]. \end{aligned} \tag{6.11}$$

- **M-step:** After finding out the Q function, we update the parameters for the next iteration, $\Theta^{(j+1)}$, which can be achieved by finding the gradient functions $g(.)$ and the Hessian matrix $H(.)$ of Q functions and then apply the Newton-Raphson algorithm to maximize the Q function for the next iteration. The parameters can be updated separately as Poisson part $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and random effect part $\boldsymbol{\phi}$.

  - For the Poisson part

$$\begin{aligned} \boldsymbol{\beta}_i^{(j+1)} &= \boldsymbol{\beta}_i^{(j)} - H^{-1}(\boldsymbol{\beta}_i^{(j)})g(\boldsymbol{\beta}_i^{(j)}), \quad i = 1,2 \\ g(\boldsymbol{\beta}_i^{(j)}) &= \mathbf{Z}_i^T\mathbf{V}_i^{(g)} \quad H(\boldsymbol{\beta}_i^{(j)}) = \mathbf{Z}_i^T\mathbf{D}_i^{(H)}\mathbf{Z}_i \\ \mathbf{V}_i^{(g)} &= \left(\left\{k_{i,t} - \lambda_{i,t}^{(j)}\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\theta_i]\right\}_{t=1,\ldots,n}\right) \\ \mathbf{D}_i^{(H)} &= \text{diag}\left(\left\{-\lambda_{i,t}^{(j)}\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[\theta_i]\right\}_{t=1,\ldots,n}\right) \end{aligned} \tag{6.12}$$

  - For the random effect part, we need to derive the first and second order derivatives of $\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and then the take posterior expectation to construct its gradient functions and Hessian matrix. In the following, we derive the derivatives for those three mixing densities defined in the last session. Different mixing densities will affect the way we calculate the posterior expectation, and in many cases, we have to rely on numerical evaluation. However, some posterior expectations can be simplified to reduce computational cost when implementing the EM algorithm in practice.

(a) **Univariate Gamma density**

This can be regarded as a special case because the posterior density is known in closed form as another univariate Gamma density with different

parameters.

$$\theta|\Theta^{(j)}, \mathbf{k}_t \sim \text{Gamma}\left(\phi^{(j)} + k_{1,t} + k_{2,t},\ \phi^{(j)} + \lambda_{1,t}^{(j)} + \lambda_{2,t}^{(j)}\right) \tag{6.13}$$

Then, the posterior expectation when updating $\boldsymbol{\beta}_i$ can be simplified as

$$\mathbb{E}_{\theta,t}^{(j)}[\theta_1] = \mathbb{E}_{\theta,t}^{(j)}[\theta_2] = \frac{\phi^{(j)} + k_{1,t} + k_{2,t}}{\phi^{(j)} + \lambda_{1,t}^{(j)} + \lambda_{2,t}^{(j)}}. \tag{6.14}$$

Finally, to update ϕ

$$\begin{aligned}
\phi^{(j+1)} &= \phi^{(j)} - \frac{g(\phi^{(j)})}{h(\phi^{(j)})}, \\
g(\phi^{(j)}) &= n(\log\phi^{(j)} - \Psi(\phi^{(j)}) + 1) + \sum_{t=1}^{n}\left(\mathbb{E}_{\theta,t}^{(j)}[\log\theta] - \mathbb{E}_{\theta,t}^{(j)}[\theta]\right) \\
h(\phi^{(j)}) &= n((\phi^{(j)})^{-1} - \Psi'(\phi^{(j)})),
\end{aligned} \tag{6.15}$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $\Psi'(x)$ are digamma and trigamma functions respectively. The posterior expectation $\mathbb{E}_{\theta,t}^{(j)}[\log\theta]$ is given by

$$\mathbb{E}_{\theta,t}^{(j)}[\log\theta] = \Psi\left(\phi^{(j)} + k_{1,t} + k_{2,t}\right) - \log\left(\phi^{(j)} + \lambda_{1,t}^{(j)} + \lambda_{2,t}^{(j)}\right) \tag{6.16}$$

(b) **Bivariate Lognormal density**

In this case, there is no analytic expression for the posterior density. However, it can be transformed in the following way

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\Theta^{(j)}, \mathbf{k}_t) &= \frac{\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_\Sigma^{LN}(\boldsymbol{\theta})}{\int_0^\infty\int_0^\infty \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_\Sigma^{LN}(\boldsymbol{\theta}) d\theta_1 d\theta_2} \\
&= \frac{\eta(e^{\boldsymbol{\epsilon}}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_\Sigma^{N}(\boldsymbol{\epsilon})}{\int_0^\infty\int_0^\infty \eta(e^{\boldsymbol{\epsilon}}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_\Sigma^{N}(\boldsymbol{\epsilon}) d\epsilon_1 d\epsilon_2} \\
&=: \pi(\boldsymbol{\epsilon}|\Theta^{(j)}, \mathbf{k}_t).
\end{aligned} \tag{6.17}$$

Then, all posterior expectations with respect to $\boldsymbol{\theta}$ can be transformed into expectations with respect to $\boldsymbol{\epsilon}$ such that $\mathbb{E}_{\boldsymbol{\theta},t}^{(j)}[h(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)}[h(e^{\boldsymbol{\epsilon}})]$. Under this transformation, all the posterior expectations can be evaluated by Gauss-Hermite quadrature. Furthermore,

$$\mathbb{E}_{\boldsymbol{\theta},t}^{(j)}[\log f_\Sigma^{LN}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)}[\log f_\Sigma^{N}(\boldsymbol{\epsilon}) - \epsilon_1 - \epsilon_2] \propto \mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)}[\log f_\Sigma^{N}(\boldsymbol{\epsilon})].$$

To update $\boldsymbol{\phi} = \{\phi_1, \phi_2, \rho\}$,

$$\begin{aligned}
\boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)}) g(\boldsymbol{\phi}^{(j)}) \\
g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{\epsilon},t} \left[\frac{\partial \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r} \right] \\
H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{\epsilon},t} \left[\frac{\partial^2 \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s} \right],
\end{aligned} \tag{6.18}$$

where the subscript r denotes the r-th element of a vector and r, s denotes r-th row s-th column entry of a matrix. The first and second order derivatives are given by

$$\begin{aligned}
\frac{\partial \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \phi_i} &= -\frac{1}{1-\rho^2}\left(-\frac{1}{\phi_i^3}\epsilon_i^2 + \frac{\rho\phi_{3-i}}{2\phi_i^2}\epsilon_i - \frac{\rho}{2\phi_{3-i}}\epsilon_{3-i} + \frac{\rho}{\phi_i^2\phi_{3-i}}\epsilon_1\epsilon_2 + \frac{\phi_i - \rho\phi_{3-i}}{4}\right) - \frac{1}{\phi_i} \\
\frac{\partial \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \rho} &= -\frac{\rho}{(1-\rho^2)^2}\left(\frac{1}{\phi_1^2}\epsilon_1^2 + \frac{1}{\phi_2^2}\epsilon_2^2 + (1 - \frac{\rho\phi_2}{\phi_1})\epsilon_1 + (1 - \frac{\rho\phi_1}{\phi_2})\epsilon_2 + \frac{\phi_1^2+\phi_2^2}{4} - \frac{\rho\phi_1\phi_2}{2}\right) \\
&\quad + \frac{1}{1-\rho^2}\left(\frac{\phi_2}{2\phi_1}\epsilon_1 + \frac{\phi_1}{2\phi_2}\epsilon_2 + \frac{1}{\phi_1\phi_2}\epsilon_1\epsilon_2 + \frac{\phi_1\phi_2}{4}\right) + \frac{\rho}{1-\rho^2} \\
\frac{\partial^2 \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \phi_i^2} &= -\frac{1}{1-\rho^2}\left(\frac{3}{\phi_i^4}\epsilon_i^2 - \frac{\rho\phi_{3-i}}{\phi_i^3}\epsilon_i - \frac{2\rho}{\phi_i^3\phi_{3-i}}\epsilon_1\epsilon_2 + \frac{1}{4}\right) + \frac{1}{\phi_i^2} \\
\frac{\partial^2 \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \rho^2} &= \frac{1+3\rho^2}{(1-\rho^2)^3}\left(\frac{1}{\phi_1^2}\epsilon_1^2 + \frac{1}{\phi_2^2}\epsilon_2^2 + (1 - \frac{\rho\phi_2}{\phi_1})\epsilon_1 + (1 - \frac{\rho\phi_1}{\phi_2})\epsilon_2\right) \\
&\quad + \frac{1+3\rho^2}{(1-\rho^2)^3}\left(-\frac{2\rho}{\phi_1\phi_2}\epsilon_1\epsilon_2 + \frac{\phi_1^2+\phi_2^2}{4} - \frac{\rho\phi_1\phi_2}{4}\right) \\
&\quad + \frac{4\rho}{(1-\rho^2)^2}\left(\frac{\phi_2}{2\phi_1}\epsilon_1 + \frac{\phi_1}{\phi_2}\epsilon_2 + \frac{1}{\phi_1\phi_2}\epsilon_1\epsilon_2 + \frac{\phi_1\phi_2}{4}\right) + \frac{1+\rho^2}{(1-\rho^2)^2} \\
\frac{\partial^2 \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \phi_1 \partial \phi_2} &= \frac{\rho}{1-\rho^2}\left(-\frac{1}{2\phi_1^2}\epsilon_1 - \frac{1}{2\phi_2^2}\epsilon_2 + \frac{1}{\phi_1^2\phi_2^2}\epsilon_1\epsilon_2 + \frac{1}{4}\right) \\
\frac{\partial^2 \log f_\Sigma^N(\boldsymbol{\epsilon})}{\partial \rho \partial \phi_i} &= -\frac{2\rho}{(1-\rho^2)^2}\left(-\frac{1}{\phi_i^3}\epsilon_i^2 + \frac{\rho\phi_{3-i}}{2\sigma_i^2}\epsilon_i - \frac{\rho}{2\phi_{3-i}}\epsilon_{3-i} + \frac{\rho}{\phi_i^2\phi_{3-i}}\epsilon_1\epsilon_2 + \frac{1}{4}(\phi_i - \rho\phi_{3-i})\right) \\
&\quad + \frac{1}{1-\rho^2}\left(-\frac{\phi_{3-i}}{2\phi_i^2}\epsilon_i + \frac{1}{2\phi_{3-i}}\epsilon_{3-i} - \frac{1}{\phi_i^2\phi_{3-i}}\epsilon_1\epsilon_2 + \frac{\phi_{3-i}}{4}\right).
\end{aligned}$$

Notice that all the derivatives are in the linear form of $\epsilon_1^2, \epsilon_2^2, \epsilon_1, \epsilon_2, \epsilon_1\epsilon_2$. Hence, we can evaluate these posterior expectations in each iteration once to avoid repeating calculations.

(c) **Gaussian copula paired with Gamma marginals**

In this case, there is no simplification either for the posterior density or for the posterior expectation. To update $\boldsymbol{\phi} = \{\phi_1, \phi_2, \rho\}$, we have almost

the same procedure as for the bivariate Lognormal case.

$$\begin{aligned} \boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)})g(\boldsymbol{\phi}^{(j)}) \\ g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{\theta},t}\left[\frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r}\right] \\ H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{\theta},t}\left[\frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s}\right], \end{aligned} \tag{6.19}$$

where the first and second order partial derivatives are given by

$$\begin{aligned} u_i &= \int_0^{\theta_i} \frac{\phi_i^{\phi_i}}{\Gamma(\phi_i)} y^{\phi_i - 1} e^{-\phi_i y} dy \\ \frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \phi_i} &= \left(-\frac{\rho^2}{1-\rho^2}\frac{\Phi^{-1}(u_i)}{f_{sn}(\Phi^{-1}(u_i))} + \frac{\rho}{1-\rho^2}\frac{\Phi^{-1}(u_{3-i})}{f_{sn}(\Phi^{-1}(u_i))}\right)\frac{\partial u_i}{\partial \phi_i} \\ &\quad + 1 + \log(\phi) - \Psi(\phi) + \log(\theta) - \theta \\ \frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \rho} &= \frac{\rho}{1-\rho^2} - \frac{\rho}{(1-\rho^2)^2}\left(\Phi^{-1}(u_1) + \Phi^{-1}(u_2)\right) + \frac{1+\rho^2}{(1-\rho^2)^2}\Phi^{-1}(u_1)\Phi^{-1}(u_2) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \phi_i^2} &= \left(-\frac{\rho^2}{1-\rho^2}\frac{1+\Phi^{-1}(u_i)^2}{f_{sn}(\Phi^{-1}(u_i))^2} + \frac{\rho}{1-\rho^2}\frac{\Phi^{-1}(u_1)\Phi^{-1}(u_2)}{f_{sn}(\Phi^{-1}(u_i))}\right)\left(\frac{\partial u_i}{\partial \phi_i}\right)^2 \\ &\quad + \frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \phi_i}\frac{\partial^2 u_i}{\partial \phi_i^2} + \frac{1}{\phi} - \Psi'(\phi_i) \\ \frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \rho^2} &= \frac{1+\rho^2}{(1-\rho^2)^2} - \frac{1+3\rho^2}{(1-\rho^2)^3}\left(\Phi^{-1}(u_1) + \Phi^{-1}(u_2)\right) + \frac{2\rho^3+6\rho}{(1-\rho^2)^3}\Phi^{-1}(u_1)\Phi^{-1}(u_2) \\ \frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \phi_1 \partial \phi_2} &= \frac{\rho}{1-\rho^2}\frac{\frac{\partial u_1}{\partial \phi_1}\frac{\partial u_2}{\partial \phi_2}}{f_{sn}(\Phi^{-1}(u_1))f_{sn}(\Phi^{-1}(u_2))} \\ \frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \rho \partial \phi_i} &= \left(-\frac{2\rho}{(1-\rho^2)^2}\frac{\Phi^{-1}(u_i)}{f_{sn}(\Phi^{-1}(u_i))} + \frac{1+\rho^2}{(1-\rho^2)^2}\frac{\Phi^{-1}(u_{3-i})}{f_{sn}(\Phi^{-1}(u_i))}\right)\frac{\partial u_i}{\partial \phi_i}. \end{aligned}$$

6.3 The bivariate mixed Poisson INAR(1) regression model

6.3.1 Model specifications

Let $\mathbf{X}$ and $\mathbf{R}$ be non-negative integer-valued random vectors in $\mathbb{R}^2$. Let $\mathbf{P}$ be a diagonal matrix in $\mathbb{R}^{2\times 2}$ with elements $p_i \in (0,1)$. The bivariate first-order integer-

valued autoregressive model (Bivariate INAR(1)) is defined as

$$\mathbf{X}_t = \mathbf{P} \circ \mathbf{X}_{t-1} + \mathbf{R}_t = \begin{bmatrix} p_1 & 0 \\ 0 & p_2 \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} R_{1,t} \\ R_{2,t} \end{bmatrix}, \tag{6.20}$$

where the thinning operator $\circ$ is the widely used binomial thinning operator such that $p_i \circ X_{i,t} = \sum_{k=1}^{X_{i,t}} U_k$ and U_k are independent identically distributed Bernoulli random variables with success probability p_i, i.e. $\mathcal{P}(U_k = 1) = p_i$. Hence $p_i \circ X_{i,t}$ is binomially distributed with size $X_{i,t}$ and success probability p_i. Then the distribution function $f_{p_i}(x, X_{i,t})$ can be easily written down as

$$f_{p_i}(k, X_{i,t}) = \binom{X_{i,t}}{k} p_i^k (1-p_i)^{X_{i,t}-k}. \tag{6.21}$$

Note that $p_i \circ X_{i,t}$ and $p_j \circ X_{j,t}, i \neq j$ are independent of each other. To adapt the heteroscedasticity arising from the data, $\mathbf{R}_t$ is bivariate mixed Poisson regression model such that $R_{i,t} \sim Po(\lambda_{i,t}\theta_i)$ defined in the last session. The joint distribution of the bivariate sequence $\mathbf{X}_{t+1}$ conditional on the last state $\mathbf{X}_t$ is given by

$$\begin{aligned} \mathcal{P}(\mathbf{X}_{t+1}|\mathbf{X}_t) &= \sum_{k_1=0}^{s_{1,t}} \sum_{k_2=0}^{s_{2,t}} f_{p_1}(k_1, X_{1,t}) f_{p_2}(k_2, X_{2,t}) f_{\mathbf{R}}(X_{1,t+1}-k_1, X_{2,t+1}-k_2) \\ f_{\mathbf{R}}(\mathbf{k}, t) &= \frac{\lambda_{1,t}^{k_1}}{k_1!} \frac{\lambda_{2,t}^{k_2}}{k_2!} \int_0^\infty \int_0^\infty \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t) f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) d\theta_1 d\theta_2 \\ s_{i,t} &= \min\{X_{i,t+1}, X_{i,t}\}, \end{aligned} \tag{6.22}$$

where $f_{\mathbf{R}}(\mathbf{k}, t)$ is a probability mass function of a bivariate mixed Poisson regression model with mixing density $f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$. Under this construction, the bivariate sequence $\mathbf{X}_t$ is correlated with each other and its correlation structure mainly depends on the correlation structure of innovation $\mathbf{R}_t$.

6.3.2 The EM algorithm

Similarly, the maximum likelihood estimation is not straightforward to apply as the log likelihood function

$$\ell(\Theta) = \sum_{t=1}^{n} \log \left(\sum_{k_1=0}^{s_{1,t}} \sum_{k_2=0}^{s_{2,t}} f_{p_1}(k_1, X_{1,t}) f_{p_2}(k_2, X_{2,t}) f_{\mathbf{R}}(X_{1,t+1}-k_1, X_{2,t+1}-k_2) \right)$$

has discrete convolution and double integrals. Then we can use similar techniques to decompose the log likelihood function as we did in section 2.

Given the observed bivariate sequence $\{\mathbf{X}_t\}_{t=1,\dots,n}$. Let $Y_{i,t} = p_i \circ X_{i,t-1}$ and $\Theta = \{p_1, p_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi}\}$ be the parameter space for this model. Suppose now we observe the latent variable $\{Y_t\}_{t=1,\dots n}$, then the log likelihood function becomes

$$\begin{aligned} \ell(\Theta|\mathbf{Y}) \propto & \sum_{t=1}^{n}\sum_{i=1}^{2}(Y_{i,t}\log p_i + (X_{i,t} - Y_{i,t})\log(1-p_i)) + \sum_{t=1}^{n}\log f_{\mathbf{R}}(\mathbf{R}_t, t) \\ & R_{i,t} = X_{i,t} - Y_{i,t}. \end{aligned} \tag{6.23}$$

Notice that there are still unobserved random variables $\boldsymbol{\theta}$ in R_t. In some of the examples we discuss in the last section, $f_{\mathbf{R}}(\mathbf{k}, t)$ may not have analytic expression and hence we would like to further break down the likelihood function. Suppose further that we observe the random effect $\{\boldsymbol{\theta}_t\}_{t=1\dots n}$, then the complete log likelihood becomes

$$\begin{aligned} \ell(\Theta|\mathbf{Y}, \boldsymbol{\theta}) \propto & \sum_{t=1}^{n}\sum_{i=1}^{2}(Y_{i,t}\log p_i + (X_{i,t} - Y_{i,t})\log(1-p_i)) \\ & + \sum_{t=1}^{n}\sum_{i=1}^{2}(R_{i,t}\log(\lambda_{i,t}) - \lambda_{i,t}\theta_{i,t}) + \sum_{t=1}^{n}\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta}). \end{aligned} \tag{6.24}$$

Define the following posterior density functions

$$\begin{aligned} \pi_1(\boldsymbol{y}|\Theta^{(j)}, \mathbf{X}_t, \mathbf{X}_{t-1}) &= \frac{f_{\mathbf{R}}(\mathbf{X}_{t-1} - \mathbf{y})\prod_{i=1}^{2} f_{p_i}(y_i, X_{i,t-1})}{\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})} \\ \pi_2(\boldsymbol{\theta}|\Theta^{(j)}, \mathbf{R}_t) &= \frac{\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{R}_t) f_{\boldsymbol{\theta}}(\boldsymbol{\phi})}{\int_0^\infty\int_0^\infty \eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{R}_t) f_{\boldsymbol{\phi}}(\boldsymbol{\theta}) d\theta_1 d\theta_2}, \end{aligned} \tag{6.25}$$

Define the posterior expectations with respect to real-value functions $h(.,.)$

$$\begin{aligned} \mathbb{E}^{(j)}_{\boldsymbol{y},t}[h(\boldsymbol{y})] &= \sum_{y_1=0}^{s_{1,t-1}}\sum_{y_2=0}^{s_{2,t-1}} h(\boldsymbol{y})\pi_1(\boldsymbol{y}|\Theta^{(j)}, \mathbf{X}_t, \mathbf{X}_{t-1}) \\ \mathbb{E}^{(j)}_{\boldsymbol{\theta},t}[h(\boldsymbol{\theta})|\mathbf{R}_t] &= \int_0^\infty\int_0^\infty h(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta}|\Theta^{(j)}, \mathbf{R}_t) d\theta_1 d\theta_2. \end{aligned} \tag{6.26}$$

- **E-step:** Evaluating the Q function $Q(\Theta; \Theta^{(j)})$ given the the parameters esti-

mated in the j-th iteration,

$$
\begin{aligned}
Q(\Theta;\Theta^{(j)}) &= \sum_{t=1}^{n}\sum_{i=1}^{2}(y_{i,t}^{(j)}\log p_i + (X_{i,t-1}-y_{i,t}^{(j)})\log(1-p_i)) \\
&\quad + \sum_{t=1}^{n}\sum_{i=1}^{2}(r_{i,t}^{(j)}\log(\lambda_{i,t}) - \lambda_{i,t}\hat{\theta}_{i,t}^{(j)}) + \sum_{t=1}^{n}\mathbb{E}_{\boldsymbol{y},t}^{(j)}[\mathbb{E}_{\boldsymbol{\theta},t}^{(j)}[\log f_{\boldsymbol{\phi}}(\boldsymbol{\theta})|\mathbf{R}_t]] \\
y_{i,t}^{(j)} &= \mathbb{E}_{\boldsymbol{y},t}^{(j)}[Y_i], \quad r_{i,t}^{(j)} = X_{i,t} - y_{i,t}^{(j)}, \quad \hat{\theta}_{i,t}^{(j)} = \mathbb{E}_{\boldsymbol{y},t}^{(j)}[\mathbb{E}_{\boldsymbol{\theta},t}^{(j)}[\theta_i|\mathbf{R}_t]].
\end{aligned}
\tag{6.27}
$$

After breaking down the log likelihood function, it is obvious that except for the log likelihood contributed by binomial distribution, the rest of the terms are almost the same as that of the Q-function of bivariate mixed Poisson regression model discussed in the last session, which means the updating procedure for $\boldsymbol{\beta}_i, \boldsymbol{\phi}$ will be exactly the same, but we need to evaluate different posterior expectations in this case.

- **M-step:** Similarly, we apply the Newton-Raphson algorithm to update the parameters. Based on the structure of $Q(\Theta;\Theta^{(j)})$, the parameters can be updated separately for binomial part $\boldsymbol{p}$, Poisson part $\boldsymbol{\beta}_i$ and random effect part $\boldsymbol{\phi}$

  - The binomial part can be updated simply as the following gradient function has a unique solution

$$
\begin{aligned}
g(p_i) &= \frac{\sum_{t=1}^{n} y_{i,t}^{(j)}}{p_i} - \frac{\sum_{t=1}^{n}(X_{i,t-1}-y_{i,t}^{(j)})}{1-p_i} = 0 \\
p_i^{(j+1)} &= \frac{\sum_{t=1}^{n} y_{i,t}^{(j)}}{\sum_{t=1}^{n} X_{i,t-1}}, \quad i = 1,2 \\
y_{i,t}^{(j)} &= \begin{cases} \frac{p_i^{(j)} X_{i,t-1}\mathcal{P}(\mathbf{X}_t-\mathbf{1}_i|\mathbf{X}_{t-1}-\mathbf{1}_i)}{\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}, & X_{i,t}\neq 0 \text{ and } X_{i,t-1}\neq 0 \\ 0, & \text{otherwise} \end{cases} \\
\mathbf{1}_1 &= (1,0)^T \quad \mathbf{1}_2 = (0,1)^T.
\end{aligned}
\tag{6.28}
$$

    See appendix 6.D for the derivation of $y_{i,t}^{(j)}$.

  - For the Poisson part, the updating equations are the same with different

posterior expectation

$$\begin{aligned}
\boldsymbol{\beta}_i^{(j+1)} &= \boldsymbol{\beta}_i^{(j)} - H^{-1}(\boldsymbol{\beta}_i^{(j)}) g(\boldsymbol{\beta}_i^{(j)}), \quad i = 1, 2 \\
g(\boldsymbol{\beta}_i^{(j)}) &= \mathbf{Z}_i^T \mathbf{V}_i^{(g)} \quad H(\boldsymbol{\beta}_i^{(j)}) = \mathbf{Z}_i^T \mathbf{D}_i^{(H)} \mathbf{Z}_i \\
\mathbf{V}_i^{(g)} &= \left(\left\{ k_{i,t} - \lambda_{i,t}^{(j)} \hat{\theta}_{i,t}^{(j)} \right\}_{t=1,\ldots,n} \right) \\
\mathbf{D}_i^{(H)} &= \operatorname{diag} \left(\left\{ -\lambda_{i,t}^{(j)} \hat{\theta}_{i,t}^{(j)} \right\}_{t=1,\ldots,n} \right).
\end{aligned} \tag{6.29}$$

Note that when the mixing density $f_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ is univariate Gamma, the posterior expectation for $\boldsymbol{\theta}$ has a simple expression

$$\hat{\theta}_t^{(j)} = \hat{\theta}_{1,t}^{(j)} = \hat{\theta}_{2,t}^{(j)} = \frac{\phi^{(j)} + r_{1,t}^{(j)} + r_{2,t}^{(j)}}{\phi^{(j)} + \lambda_{1,t}^{(j)} + \lambda_{2,t}^{(j)}}.$$

– Similarly, for the random effect part $\boldsymbol{\phi}$,

(a) **Univariate Gamma density**

$$\begin{aligned}
\phi^{(j+1)} &= \phi^{(j)} - \frac{g(\phi^{(j)})}{h(\phi^{(j)})}, \\
g(\phi^{(j)}) &= n(\log \phi^{(j)} - \Psi(\phi^{(j)}) + 1) + \sum_{t=1}^{n} \left(\mathbb{E}_{\boldsymbol{y},t}^{(j)} [\mathbb{E}_{\theta,t}^{(j)} [\log \theta | \mathbf{R}_t]] - \hat{\theta}_t^{(j)} \right) \\
h(\phi^{(j)}) &= n((\phi^{(j)})^{-1} - \Psi'(\phi^{(j)})),
\end{aligned} \tag{6.30}$$

(b) **Bivariate Lognormal**

$$\begin{aligned}
\boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)}) g(\boldsymbol{\phi}^{(j)}) \\
g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)} \left[\frac{\partial \log f_{\Sigma}^{N}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r} | \mathbf{R}_t \right] \right] \\
H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{y},t}^{(j)} \left[\mathbb{E}_{\boldsymbol{\epsilon},t}^{(j)} \left[\frac{\partial^2 \log f_{\Sigma}^{N}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s} | \mathbf{R}_t \right] \right],
\end{aligned} \tag{6.31}$$

(c) **Gaussian copula paired with Gamma marginals**

$$\begin{aligned}
\boldsymbol{\phi}^{(j+1)} &= \boldsymbol{\phi}^{(j)} - H^{-1}(\boldsymbol{\phi}^{(j)})g(\boldsymbol{\phi}^{(j)}) \\
g(\boldsymbol{\phi}^{(j)})_r &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{y},t}\left[\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}\left[\frac{\partial \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r}|\mathbf{R}_t\right]\right] \\
H(\boldsymbol{\phi}^{(j)})_{r,s} &= \sum_{t=1}^{n} \mathbb{E}^{(j)}_{\boldsymbol{y},t}\left[\mathbb{E}^{(j)}_{\boldsymbol{\theta},t}\left[\frac{\partial^2 \log f_{GC}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}_r \partial \boldsymbol{\phi}_s}|\mathbf{R}_t\right]\right].
\end{aligned} \tag{6.32}$$

The partial derivatives inside expectations are derived in the last section.

Remark This model as well as the EM algorithm can be extent to multivariate case straightforwardly. All the steps and the general form of the formula of the EM algorithm in the multivariate case are exactly the same. The only problem is that it would become cumbersome to evaluate the transition probability $\mathcal{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$ as dimension of $\mathbf{X}_t$ increases.

6.4 Empirical analysis

6.4.1 Data description and model fitting

The data used in this section come from the Local Government Property Insurance Fund (LGPIF) from the state of Wisconsin. On previous application on this dataset, interested reader can refer to Frees et al. (2016), Lee and Shi (2019) and Jeong et al. (2023). This fund provides property insurance to different types of government units, which includes villages, cities, counties, towns and schools. The LGPIF contains three major groups of property insurance coverage, namely building and contents (BC), contractors' equipment (IM) and motor vehicles (PN, PO, CN, CO). For exploratory purposes, we focus on modelling jointly the claim frequency of IM, denoted as X_1, and comprehensive new vehicles collision (CN), denoted as X_2 as they are both related to land transport. The insurance data cover the period over 2006 - 2010 with 1234 policyholders in total. Only $n_1 = 1048$ of them have complete data over the period 2006-2010, which will be the training dataset. The last year 2011 with $n_2 = 1025$ policyholders, which is the same set of policyholders as in the training dataset, out of 1098 policyholders will be the test dataset. Denote the IM type and CN type claim frequency for a particular policyholder as $X^{(h)}_{1,t}, X^{(h)}_{2,t}$ respectively, where h is the identifier for each policyholder and t is the year. Then the relationship between $X_{i,t}$ and $X^{(h)}_{i,t}$ is simply $X_{i,t} = \sum_h X^{(h)}_{i,t}$ with $i = 1, 2$.

Some basic statistical analysis is shown in the following Table 6.1 and Figure 6.1. The proportion of zeros for two types of claims are all over 90% during 2006-2010. Both types of claim shows overdispersion as the variance are all higher than their mean over years and the overdispersion for $X_{2,t}$ is even stronger than that of $X_{1,t}$, which indicate the need to apply overdispersed distribution model for the data. The correlation tests over years imply that it is reasonable to introduce correlation structure between $X_{1,t}$ and $X_{2,t}$. The proportion of zeros and kurtosis show that the marginal distributions of $X_{1,t}, X_{2,t}$ are all positively skewed and exhibit a fat-tailed structure which indicates the appropriateness of adopting a positive skewed and fat-tailed distribution (Log Normal distribution). Last but not least, the correlation tests illustrated in Table 6.2 do support the appropriateness of introduction of time series term in modelling the claim sequence.

Table 6.1: Summary statistics of two types of claims over years. The correlations test is a one-sided test where the alternative hypothesis is "The sample correlation is greater than 0"

	2006	2007	2008	2009	2010
Proportion of zeros $X_{1,t}$	0.9685	0.9542	0.9552	0.9504	0.9590
Proportion of zeros $X_{2,t}$	0.9342	0.9332	0.9399	0.9370	0.9323
Kurtosis of $X_{1,t}$	85.7500	86.6479	41.8491	43.0183	126.6879
Kurtosis of $X_{2,t}$	53.9183	61.7740	111.2810	184.1395	133.9228
P-value of correlation test	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6.2: Correlation test for $X_{i,t}$ and $X_{i,t-1}$. The test is a one-sided test where the alternative hypothesis is "The sample correlation is greater than 0"

	X_1	X_2
correlation	0.4062	0.7478
p-value	0.0000	0.0000

The description and some summary statistics for all the explanatory variables (covariates $z_{1,t}, z_{2,t}$) that are relevant to $X_{1,t}, X_{2,t}$ are shown in Table 6.3. Variables 1- 5 including 'TypeVillage' are categorical variables to indicate the entity types of a policyholder. Due to the strongly heavy-tailed structure appearing in variables 6 and 9, which can drastically distort the model fitting, those variables are transformed by means of the 'rank' function in R software and then standardized, which can mitigate the effect of outliers. Variables 6-8 are relevant to IM claim $X_{1,t}$, while variables 9,10 provide information for CN claims $X_{2,t}$. The covariate $z_{1,t}$ includes variables 1-8, and $z_{2,t}$ contains variables 1-5 and variables 9,10. These covariates act as the regression part for $\lambda_{i,t}$ mentioned in section 2, which may help to explain part of the heterogeneity within $X_{1,t}$ and $X_{2,t}$.

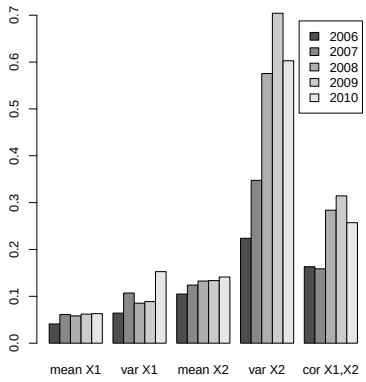


Figure 6.1: Summary statistics (mean, variance and correlation) for each type of claims across all the policyholders over the years.

Table 6.3: summary statistics for the explanatory variables

Variable index	Variable name	Type	Description	Proportion/Mean
1	TypeCity	Categorical	Indicator for city entity	0.1400
2	TypeCounty	Categorical	Indicator for county entity	0.0578
3	TypeMisc	Categorical	Indicator for miscellaneous entity	0.1104
4	TypeSchool	Categorical	Indicator for school entity	0.2817
5	TypeTown	Categorical	Indicator for town entity	0.1728
-	TypeVillage	Categorical	Indicator for village entity (reference category)	0.2373
6	CoverageIM	Continuous	Coverage amount of IM(transformed)	0
7	InDeductIM	Continuous	Log deductible amount for inland marine	5.3400
8	NoClaimCreditIM	Binary	Indicator for no IM claims in prior year	0.4210
9	CoverageCN	Countinuous	Coverage amount of CN (transformed)	0
10	NoClaimCreditCN	Binary	Indicator for no CN claims in prior year	0.0897

Due to the large computational cost for evaluating the partial derivatives of copula case (large sample size), all the models except the copula case discussed in Section 2 and Section 3 are applied to model the joint behaviour of $X_{1,t}^{(h)}, X_{2,t}^{(h)}$ across all the policyholders. Instead, a simulation study in the appendix 6.E shows that the EM algorithm does work for copula case.

Since we would like to model the whole behaviour rather than the individual one, the the likelihood function would simply become

$$\ell(\Theta) = \sum_{h=1}^{n_1} \ell_h(\Theta) = \sum_{h=1}^{n_1} \sum_{t=1}^{4} \log \Pr(X_{1,t+1}^{(h)}, X_{2,t+1}^{(h)} | X_{1,t}^{(h)}, X_{2,t}^{(h)}), \tag{6.33}$$

where $\ell_h(\Theta)$ is the log likelihood function for policyholder h. Note that all the policyholders with the same type of claim $X_{i,}$ will share the same set of parameters $\{p_1, p_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi}\}$. In addition, it is necessary to show the appropriateness of introducing crosscorrelation and autocorrelation in BINAR(1) model. Then we also fit the data to following models.

1. The joint distribution of $X_{1,t}^{(h)}$ and $X_{2,t}^{(h)}$ are characterized by two independent mixed Poisson (TMP)

$$X_1^{(h)} \sim Pois(\lambda_1\theta_1), \quad X_2^{(h)} \sim Pois(\lambda_2\theta_2), \tag{6.34}$$

   where θ_1 and θ_2 are independent random variables, either Gamma or Log Normal.

2. The joint distribution of $X_{1,t}^{(h)}$ and $X_{2,t}^{(h)}$ are assumed to be bivariate mixed Poisson distribution (BMP) with different probability mass function according to the choice of mixing densities, see equations (6.2) and (6.4).

3. The joint distribution of $X_{1,t}^{(h)}$ and $X_{2,t}^{(h)}$ are characterized by two independent INAR(1) models (TINAR)

$$\begin{aligned} X_{1,t}^{(h)} &= p_1 \circ X_{1,t-1}^{(h)} + R_{1,t} \\ X_{2,t}^{(h)} &= p_2 \circ X_{2,t-1}^{(h)} + R_{2,t}, \end{aligned}$$

   where $R_{i,t} \sim Pois(\lambda_{i,t}\theta_{i,t}), i = 1, 2$ and random effect $\theta_{i,t}$ is independent of i and t.

For comparison purpose, we fit these univariate and bivariate Poisson mixture models with training dataset starting from 2007 because they do not need to consider the lag responses. When it comes to the initial values, we use the following. Lag one correlation of each sequence serves as the initial value of p_i. We fit a Poisson generalized linear model for each sequence to obtain the initial values of $\boldsymbol{\beta}_i$. Finally, we used the moment estimates of the bivariate Poisson mixture model (without regression) for initial values of $\boldsymbol{\phi} = \{\phi_1, \phi_2, \rho\}$. All the estimation is performed in R software where we implement the EM algorithms derived in previous sections. The standard deviations of the estimators are calculated by inverting the observed information of matrix from the incomplete log-likelihood function (the log likelihood function without unobserved latent variables).

Model fitting results are shown in Table 6.4. Within the same class of models, compared to univariate Gamma as mixing density, the Log Normal case allows more flexible structures to capture different distributional behaviour within two types of claims. Hence we can observe the improvement of AIC from univariate Gamma case to Log Normal case and hence it is no surprise that the Log Normal is always the best choice within the same class of model. Among different classes of models, it is clear that the adoption of autocorrelation component significantly improves the

model fitting. Finally, the significant improvement in terms of AIC from TINAR to BINAR, as well as from TMP to BMP, indicates that it is appropriate to introduce cross correlation between two sequences X_1, X_2. The estimated parameters via EM algorithm are shown in Table 6.5.

Table 6.4: Goodness of fit for different models with different choices of mixing densities. For the class TMP and TINAR, row and column stand for mixing density of θ_1, θ_2, respectively. The grey cells indicate the best one within the same class of models.

	Mixing density	Gamma AIC | BIC	Log Normal AIC | BIC
TMP	Gamma	3073.332 | 3193.810	3067.119 | 3187.596
	Log Normal	3072.287 | 3192.765	3066.074 | 3186.551
TINAR	Gamma	2999.957 | 3133.117	2999.843 | 3133.003
	Log Normal	2998.326 | 3131.485	2998.212 | 3131.371
BMP	/	3073.176 | 3191.329	3055.066 | 3181.885
BINAR	/	2996.291 | 3123.109	2990.744 | 3130.245

6.4.2 Predictive performance

In insurance claims modelling, it is more useful to check the overall distribution for all policyholders rather than prediction of the claim frequency for each policyholder, which can be used for premium calculation, risk management, and so forth. To evaluate the predictive performance, we then calculate the predicted claim frequencies $\text{Freq}(\mathbf{X}_t|\mathbf{X}_{t-1}, \hat{\Theta})$, which are the sum of individual probabilities $\mathcal{P}(\mathbf{X}_t^{(h)}|\mathbf{X}_{t-1}^{(h)}, \hat{\Theta})$ of joint events $(X_1^{(h)}, X_2^{(h)}) \in \{(i,j), 0 \leqslant i,j \leqslant 10\}$ based on the estimated parameters, and compare these to the observed frequencies from the test sample $(X_{1,2011}, X_{2,2011})$ (year 2011). In addition to our proposed BINAR model with Log Normal mixing density, we also compute predictive performance of the best TMP, TINAR, BMP models from Table 6.4 as the benchmark for comparison purposes. Based on the predictive claim frequencies, one can also compute expected number of claims marginally $(\mathbb{E}[X_1], \mathbb{E}[X_2])$,

$$\begin{aligned} \mathbb{E}[X_1] &= \sum_{k_1=0}^{10} \sum_{k_2=0}^{10} k_1 \text{Freq}(k_1, k_2|\mathbf{X}_{2010}, \hat{\Theta}) \\ \mathbb{E}[X_2] &= \sum_{k_1=0}^{10} \sum_{k_2=0}^{10} k_3 \text{Freq}(k_1, k_2|\mathbf{X}_{2010}, \hat{\Theta}) \end{aligned} \tag{6.35}$$

Table 6.5: Parameter estimation for the Bivariate mixed Poisson regression model and the bivariate INAR model of insurance claim frequency data with EM algorithm. For each entry, the upper one is the estimate and the estimated standard deviations are indicated in round brackets.

Estimate \ Model	BMP				BINAR			
	BPGA		BPLN		BINAR(1)-GA		BINAR(1)-LN	
	$X_{1,t}$	$X_{2,t}$	$X_{1,t}$	$X_{2,t}$	$X_{1,t}$	$X_{2,t}$	$X_{1,t}$	$X_{2,t}$
p_i					0.1238	0.2904	0.1118	0.2761
					(0.0373)	(0.0378)	(0.0384)	(0.03950)
ϕ_i	1.1742		1.0668	0.8408	0.9495		1.0559	0.9403
	(0.1656)		(0.1167)	(0.0696)	(0.1662)		(0.1274)	(0.0931)
ρ			0.5895				0.6063	
			(0.1485)				(0.1782)	
Intercept	-3.7300	-5.4277	-3.6811	-5.4836	-3.7980	-5.9744	-3.7966	-6.0363
	(0.4835)	(0.3612)	(0.5175)	(0.3743)	(0.5158)	(0.4428)	(0.5412)	(0.4607)
TypeCity	-0.2183	0.5354	-0.2830	0.5234	-0.2242	0.6673	-0.2735	0.6625
	(0.2405)	(0.2259)	(0.2565)	(0.2306)	(0.2823)	(0.2577)	(0.2657)	(0.2888)
TypeCounty	0.5269	1.2034	0.6190	1.1312	0.5682	1.3290	0.6200	1.2744
	(0.2605)	(0.2074)	(0.2809)	(0.2119)	(0.2811)	(0.2643)	(0.2930)	(0.2698)
TypeMisc	-2.0565	-0.4822	-2.0968	-0.4916	-2.0110	-0.1141	-2.0360	-0.1279
	(1.0192)	(0.6349)	(1.0251)	(0.6355)	(1.0210)	(0.6567)	(1.0253)	(0.6604)
TypeSchool	0.0678	-0.0068	-0.0768	-0.0190	-0.0387	0.1559	-0.0945	0.1367
	(0.3242)	(0.2291)	(0.3401)	(0.2318)	(0.3587)	(0.2811)	(0.36074)	(0.2852)
TypeTown	-0.4060	-1.1058	-0.4427	-1.1294	-0.3565	-0.8941	-0.3830	-0.9257
	(0.2992)	(0.4289)	(0.3064)	(0.4309)	(0.3037)	(0.4794)	(0.3091)	(0.4832)
CoverageIM	1.5299		1.4519		1.4543		1.4219	
	(0.1971)		(0.2062)		(0.2126)		(0.2163)	
InDeductIM	0.0241		0.0274		0.0170		0.0233	
	(0.0730)		(0.0799)		(0.0788)		(0.0835)	
NoClaimCreaditIM	-0.6557		–0.6053		-0.4569		-0.4382	
	(0.1421)		(0.1552)		(0.1570)		(0.1665)	
CoverageCN		2.3947		2.4422		2.4227		2.4818
		(0.1819)		(0.1903)		(0.2210)		(0.2210)
NoClaimCreaditCN		-0.6752		-0.6786		-0.3047		-0.3139
		(0.1652)		(0.1663)		(0.1811)		(0.1811)

and measure the Predictive Sum of Square error:

$$\text{PSSE} = (\mathbb{E}[X_1] - X_{1,2011})^2 + (\mathbb{E}[X_2] - X_{2,2011})^2 \tag{6.36}$$

On the other hand, the log likelihood on test samples (TLL) can also be a measure of predictive performance for each model.

$$TLL = \sum_{h}^{n_2} \log \mathcal{P}(\mathbf{X}_t^{(h)} | \mathbf{X}_{t-1}^{(h)}, \hat{\Theta}). \tag{6.37}$$

All the results are summarised in Table 6.6 and 6.7 and it is clear that our proposed model, bivariate INAR(1), has the best predictive performance with the smallest

PSSE among all other models. Furthermore, TLL result shows that the bivariate INAR outperforms all other models, which is consistent with the model fitting result in Table 6.4.

Table 6.6: Predictive and observed joint frequencies for each models

	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)
TMP	901.40	36.53	11.95	38.30	4.03	2.20	5.37	0.97	0.62
TINAR	903.99	37.47	11.70	37.42	4.16	2.30	4.87	0.93	0.61
BMP	904.09	36.87	11.75	36.91	4.09	2.41	4.81	0.81	0.63
BINAR	906.03	37.67	11.55	36.31	4.16	2.45	4.45	0.79	0.61
Observed	940	26	6	20	6	2	4	1	1

Table 6.7: predictive Marginal claim frequency

	TMP	TINAR	BMP	BINAR	Observed
$\mathbb{E}[X_1]$	79.18	74.17	77.06	72.55	78
$\mathbb{E}[X_2]$	158.89	153.28	157.34	152.68	117
PSSE	1756.164	1330.907	1628.199	1302.765	/
TLL	-348.3537	-340.4619	-345.9555	-338.4115	/

6.4.3 Application to ratemaking

In this subsection, the analysis of best fitted models from Table 6.4 for ratemaking is conducted. We select three representative risk profiles under different models, named Good, Average and Bad, illustrated in Table 6.8. These three risk profiles are selected according to the sign and size of the coefficients in Table 6.5 and those variables are not mentioned in the following table are taken to be 0. Note that CoverageIM and CoverageCN are selected according to their empirical distribution on test data.

Table 6.8: The risk profiles for two claim sequences

		TypeCounty	TypeMisc	CoverageIM
	Good	0	1	-1.43
$\mathbf{Z}_1$	Average	1	0	0
	Bad	1	0	1
		TypeCounty	TypeMisc	CoverageCN
	Good	0	1	-0.56
$\mathbf{Z}_2$	Average	1	0	1.6
	Bad	1	0	2.15

We then evaluate the mean and variance of $X_{1,t}^{(h)} + X_{2,t}^{(h)}$ under each best TMP, TINAR, BMP and BINAR according to Table 6.4. The mean and variance for one policyholder is given by two quantity, i.e. $\mathbb{E}[X_1^{(h)} + X_2^{(h)}|\hat{\Theta}, \mathbf{X}_{t-1}, \mathbf{Z}_1, \mathbf{Z}_2]$ and $\mathrm{Var}(X_1^{(h)} + X_2^{(h)}|\hat{\Theta}, \mathbf{X}_{t-1}, \mathbf{Z}_1, \mathbf{Z}_2)$. They have following explicit formulae

$$\begin{aligned}
\mathbb{E}[R_{1,t} + R_{2,t}|\hat{\Theta}, \mathbf{Z}_1, \mathbf{Z}_2] &= \hat{\lambda}_1 + \hat{\lambda}_2 = e^{\mathbf{Z}_1\hat{\beta}_1} + e^{\mathbf{Z}_2\hat{\beta}_2} \\
\mathrm{Var}(R_{1,t} + R_{2,t}|\hat{\Theta}, \mathbf{Z}_1, \mathbf{Z}_2) &= \hat{\lambda}_1^2(e^{\hat{\sigma}_1} - 1) + \hat{\lambda}_2^2(e^{\hat{\sigma}_2} - 1) + \hat{\lambda}_1 + \hat{\lambda}_2 \\
&\quad + 2\hat{\lambda}_1\hat{\lambda}_2(e^{\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2} - 1) \\
\mathbb{E}[X_1^{(h)} + X_2^{(h)}|\hat{\Theta}, \mathbf{X}_{t-1}, \mathbf{Z}_1, \mathbf{Z}_2] &= \hat{p_1}X_{1,t-1}^{(h)} + \hat{p_2}X_{2,t-1}^{(h)} \\
&\quad + \mathbb{E}[R_{1,t} + R_{2,t}|\hat{\Theta}, \mathbf{Z}_1, \mathbf{Z}_2] \\
\mathrm{Var}(X_1^{(h)} + X_2^{(h)}|\hat{\Theta}, \mathbf{X}_{t-1}, \mathbf{Z}_1, \mathbf{Z}_2) &= \hat{p_1}(1-\hat{p_1})X_{1,t-1}^{(h)} + \hat{p_2}(1-\hat{p_2})X_{2,t-1}^{(h)} \\
&\quad + \mathrm{Var}(R_{1,t} + R_{2,t}|\hat{\Theta}, \mathbf{Z}_1, \mathbf{Z}_2)
\end{aligned} \tag{6.38}$$

Table 6.9 and 6.10 summarised the mean and variance under different risk profiles and different claim history structure $(X_{1,t-1}^{(h)}, X_{2,t-1}^{(h)})$. As TMP and BMP do not depend on claim history, their mean and variance are all the same within the same risk profiles. It is interesting to see that the variance of INAR models are smaller than that of mixed Poisson models in many cases.

Table 6.9: Premium calculations from different models: Means.

Profile	$(X_{1,t-1}^{(h)}, X_{2,t-1}^{(h)})$	TMP	TINAR	BMP	BINAR
Good	(0,0)	0.0010	0.0009	0.0010	0.0009
	(0,1)	0.0010	0.2856	0.0010	0.2771
	(1,0)	0.0010	0.1194	0.0010	0.1128
	(1,1)	0.0010	0.4040	0.0010	0.3889
Average	(0,0)	0.6361	0.4690	0.6877	0.4951
	(0,1)	0.6361	0.7537	0.6877	0.7713
	(1,0)	0.6361	0.5875	0.6877	0.6070
	(1,1)	0.6361	0.8722	0.6877	0.8832
Bad	(0,0)	2.4873	1.8631	2.6553	1.9484
	(0,1)	2.4873	2.1478	2.6553	2.2246
	(1,0)	2.4873	1.9816	2.6553	2.0603
	(1,1)	2.4873	2.2663	2.6553	2.3364

Table 6.10: Premium calculations from different models: Variances.

Profile	$(X_{1,t-1}^{(h)}, X_{2,t-1}^{(h)})$	TMP	TINAR	BMP	BINAR
Good	(0,0)	0.0010	0.0009	0.0010	0.0009
	(0,1)	0.0010	0.2045	0.0010	0.2008
	(1,0)	0.0010	0.1053	0.0010	0.1003
	(1,1)	0.0010	0.3090	0.0010	0.3001
Average	(0,0)	0.8746	0.6869	1.1563	0.8221
	(0,1)	0.8746	0.8905	1.1563	1.0220
	(1,0)	0.8746	0.7914	1.1563	0.9215
	(1,1)	0.8746	0.9950	1.1563	1.1214
Bad	(0,0)	6.0635	5.2554	9.6217	6.9969
	(0,1)	6.0635	5.4591	9.6217	7.1968
	(1,0)	6.0635	5.3599	9.6217	7.0962
	(1,1)	6.0635	5.5635	9.6217	7.2961

6.5 Concluding remarks

In this paper, we consider a new family of bivariate mixed Poisson INAR(1) regression models for modelling multiple time series of different types of claim counts. The proposed family of models accounts for bivariate overdispersion and, similarly to copula-based models, allows for interactions of different signs and magnitude among the two count response variables without using the finite differences of the copula representation which may result in numerical instability in the ML estimation procedure. For illustrative purposes, we derived the BINAR(1)-LN and BINAR(1)-GGA regression models which can be regarded as competitive alternatives to the BINAR(1)-GA regression model for modelling time series of count data. Furthermore, from a computational statistics standpoint, the EM type algorithms we developed for ML estimation of the parameters of all the models were easily implementable and were shown to perform well when we exemplified our approach on LGPIF data from the state of Wisconsin. At this point, it should be noted that we considered the bivariate case and the Gamma and Lognormal correlated random effects for expository purposes. Moreover, the EM estimation framework we proposed is sufficiently flexible and can be used for other continuous mixing densities with a unit mean and, unlike copula-based models, which also allow for both positive and negative correlations, generalizations to any vector size response variables are straightforward. However, in the latter case, EM estimation may be chronologically demanding due to algebraic intractability. Nevertheless, in such cases, due to the structure of the EM algorithm for multivariate INAR(1) models with correlated ran-

dom effects, the E- and M-steps can be executed in parallel across multiple threads to exploit the processing power available in multicore machines.

Finally, an interesting topic for further research would be to also take into account cross autocorrelation, proceeding along similar lines as in Bermúdez et al. (2018).

6.A abbreviations

Here is a table for all the abbreviations used in this paper

Table 6.11: The explanation of the Abbreviations used in

Abbreviation	Interpretation
BP...	Bivariate Mixed Poisson regression model...
BPGA	∼with univarite Gamma as mixing density
BPLN	∼with bivariate log normal as mixing density
BPGGA	∼with bivarite Gaussian Copula paried with univariate Gamma as mixing density
BINAR(1) - ...	Bivariate Integer-valued autoregssive model ...
BINAR(1) - GA	∼with BPGA as innovations
BINAR(1) - LN	∼with BPLN as innovations
BINAR(1) - GGA	∼with BPGGA as innovations

6.B The Gauss-Hermite quadrature in the high dimensional setting

In this session, we introduce how to transform an integral with respect to multivariate normal density function into a multi-dimension Gauss-Hermite quadrature rule. Starting from one dimensional case, the way we calculate the following integral

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx,$$

where $X \sim N(\mu, \sigma^2)$, is first to make a linear transformation of integrand and then apply the quadrature rule directly:

$$\begin{aligned}\mathbb{E}[h(x)] &= \int_{-\infty}^{\infty} h(\sqrt{2}\sigma y + \mu) \frac{1}{\sqrt{\pi}} \exp\left\{-y^2\right\} dy \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{n} h(\sqrt{\sigma}\xi_i + \mu) w_i\end{aligned}$$

where ξ_i are the roots of Hermite polynomial of degree n, with a certain weight w_i. The quadrature rule approximation of integral will be accurate only when the function h can be well-approximated by a polynomial of degree $2n-1$ or less. Those values can be found from the R function **gauss.quad** in the package **statmod**. The idea to extend the result to high dimensional setting is straightforward. Specifically,

we need to first transform the density function into the form $\exp\{\mathbf{y}^T\mathbf{y}\}$, where $\mathbf{y} \in \mathbb{R}^{k\times 1}$, then the k-dimensional integral reduces to a k-fold Gauss-Hermit integral. Suppose the k -dimensional random vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^{k\times 1}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$. Then a linear transform for this random vector is through eigen decomposition of $\boldsymbol{\Sigma}$ such that

$$\mathbf{x} = \sqrt{2}\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} + \boldsymbol{\mu},$$

where $\mathbf{Q} = (\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_k)$ is the matrix formed by eigen vectors and $\boldsymbol{\Lambda}$ is the diagonal matrix with eigen values $(\lambda_1, \ldots, \lambda_k)$ such that $\boldsymbol{\Sigma}\boldsymbol{\nu}_i = \lambda_i\boldsymbol{\nu}_i, \quad i = 1, \ldots., k$. Then the exponent of multivariate normal density becomes

$$\begin{aligned}
&\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\\
=&\frac{1}{2}\left(\sqrt{2}\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}+\boldsymbol{\mu}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\sqrt{2}\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}+\boldsymbol{\mu}-\boldsymbol{\mu}\right)\\
=&\left(\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}\right)^T\left(\mathbf{Q}\boldsymbol{\Lambda}^{-1}\mathbf{Q}^{-1}\right)\left(\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}\right)\\
=&\mathbf{y}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\mathbf{Q}^T\mathbf{Q}\right)\boldsymbol{\Lambda}^{-1}\left(\mathbf{Q}^{-1}\mathbf{Q}\right)\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}\\
=&\mathbf{y}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}\\
=&\mathbf{y}^T\mathbf{y}.
\end{aligned}$$

Since $\boldsymbol{\Sigma}$ is symmetric, then $\mathbf{Q}^T = \mathbf{Q}^{-1}$. Finally, the k-dimensional integral becomes

$$\begin{aligned}
\mathbb{E}[h(\mathbf{X})] &= \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} h(\mathbf{x})\frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}|}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\}dx_1\ldots dx_k\\
&= \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} h\left(\sqrt{2}\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}+\boldsymbol{\mu}\right)\pi^{-\frac{k}{2}}\exp\left\{-\mathbf{y}^T\mathbf{y}\right\}dy_1\ldots dy_k\\
&\approx \pi^{-\frac{k}{2}}\sum_{i_1=1}^{n}\ldots\sum_{i_k=1}^{n} h\left(\sqrt{2}\mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\xi}+\boldsymbol{\mu}\right)w_{i_1}\ldots w_{i_k}
\end{aligned}$$

6.C The Gauss-Legendre Quadrature in the high dimensional setting

The the extension of the Gauss-Legendre quadrature rule into high dimensional situation is much more straightforward. The following m-dimensional integral can

be approximated by

$$\int_{a_1}^{b_1} \dots \int_{a_m}^{b_m} h(y_1, \dots, y_n) dy_1 \dots dy_n$$
$$\approx \left(\prod_{i=1}^{m} \frac{b_i - a_i}{2} \right) \sum_{i_1=1}^{n} \dots \sum_{i_m=1}^{n} h\left(\frac{b_1 - a_1}{2} \xi_{i_1} + \frac{b_1 + a_1}{2}, \dots, \frac{b_n - a_n}{2} \xi_{i_m} + \frac{b_n + a_n}{2} \right) w_{i_1} \dots w_{i_m}$$

where $\xi_\cdot$ are roots of Legendre polynomials of degree n and $w_\cdot$ are the corresponding weights. These can also be found easily in R by the 'gauss.quad' function in the package 'statmod'. Similarly, the h function should be well-approximated by a polynomial of degree $2n-1$ or less to ensure accuracy of the approximation.

6.D Derivation of conditional expectation

The conditional expectation $\mathbb{E}_{y,t}^{(j)}[Y_i]$ can be derived explicitly as follows. For simplicity, we just write p_1, p_2 instead of $p_1^{(j)}, p_2^{(j)}$

$$\begin{aligned}
\mathbb{E}_{y,t}^{(j)}[Y_1] &= \frac{1}{\mathcal{P}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})} \sum_{k_1=0}^{s_{1,t}} \sum_{k_2=0}^{s_{2,t}} k_1 \binom{X_{1,t-1}}{k_1} p_1^{k_1} (1-p_1)^{X_{1,t-1}-k_1} \\
&\quad \times f_{p_2}(y_2, X_{2,t-1}) f_{\boldsymbol{R}}(X_{1,t} - k_1, X_{2,t} - k_2) \\
&= \frac{p_1 X_{1,t-1}}{\mathcal{P}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})} \sum_{k_1=1}^{s_{1,t}} \sum_{k_2=0}^{s_{2,t}} \binom{X_{1,t-1}-1}{k_1-1} p_1^{k_1-1} (1-p_1)^{X_{1,t-1}-k_1} \\
&\quad \times f_{p_2}(y_2, X_{2,t-1}) f_{\boldsymbol{R}}(X_{1,t} - k_1, X_{2,t} - k_2) \\
&= \frac{p_1 X_{1,t-1}}{\mathcal{P}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})} \sum_{k_1'=0}^{s'_{1,t}} \sum_{k_2=0}^{s_{2,t}} \binom{X_{1,t-1}-1}{k_1'} p_1^{k_1'} (1-p_1)^{X_{1,t-1}-1-k_1'} \\
&\quad \times f_{p_2}(y_2, X_{2,t-1}) f_{\boldsymbol{R}}(X_{1,t} - 1 - k_1', X_{2,t} - k_2) \\
&= p_1 X_{1,t-1} \frac{\mathcal{P}(\boldsymbol{X}_t - \boldsymbol{1}_1|\boldsymbol{X}_{t-1} - \boldsymbol{1}_1)}{\mathcal{P}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})}
\end{aligned}$$

6.E Simulation study for Gaussian copula paired with Gamma marginals

This is to verify that the EM algorithms work for both the bivariate Poisson mixture model and the bivariate INAR model when the random effect is characterized by copula. Two random samples of size 500 are generated from these two models separately with pre-specified parameters, which are listed in the following table

Table 6.12: Parameter setting for simulation

parameters	$\boldsymbol{p}$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	ϕ_1, ϕ_2	ρ	$\mathbf{Z}_1$	$\mathbf{Z}_2$
values	0.4,0.5	-2,0.8,0.5	-1.5,0.5,0.3	2,3	0.5	MVN($\boldsymbol{\mu}_1, \mathbf{D}$)	MVN($\boldsymbol{\mu}_2, \mathbf{D}$)

where $\boldsymbol{\mu}_1 = (1, 0.3, 0.5)^T$, $\boldsymbol{\mu}_2 = (1, 0.2, 0.4)^T$, $\mathbf{D} = \text{diag}\{0, 1, 1\}$ and MVN stands for multivariate normal distribution. Then each model is fitted by two methods: maximising the incomplete likelihood and EM algorithms. These two methods should give almost the same results for $\boldsymbol{p}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ which determine the mean of the model and hence have a relatively large contribution to likelihood. On the other hand, this may not be the case for other parameters ϕ_1, ϕ_2, ρ which determine the variation and correlation of the model, and only contribute relative small part of the likelihood. The final log likelihood values would normally be larger than the log likelihood value evaluated at pre-specific parameters. The estimated results are given in Table 6.13

The difference between estimated parameters ϕ_1, ϕ_2, ρ and their true values seems larger than others. This is reasonable because these parameters control the variation of distribution and the log likelihood would be less sensitive to them.

Table 6.13: Simulation study for two estimation methods, where OPTIM is the optimization function implemented in R with conjugate gradient (CG) as method

	parameters	$\boldsymbol{p}$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	ϕ_1, ϕ_2	ρ	log likelihood
BPGGA	TRUE		-2 0.8 0.5	-1.5 0.5 0.3	2 3	0.5	-674.2437
	CG		-1.9428 0.8328 0.4920	-1.5659 0.4712 0.2778	3.2321 3.083	0.5666	-672.4953
	EM		-1.9554 0.8337 0.5019	-1.5662 0.4727 0.2782	3.4162 3.1606	0.6444	-672.44
BINAR(1)-GGA	TRUE	0.4 0.5	-2 0.8 0.5	-1.5 0.5 0.3	2 3	0.5	-825.1105
	CG	0.4022 0.5253	-2.0019 0.7560 0.4778	-1.7251 0.4788 0.3315	1.8934 1.2583	0.3665	-821.6101
	EM	0.4014 0.5232	-2.0028 0.7587 0.4793	-1.7198 0.4767 0.3322	2.1990 1.5932	0.4004	-821.7555

CHAPTER 7

Paper D. Cluster point processes and Poisson thinning INARMA

Abstract

In this paper, we consider Poisson thinning Integer-valued time series models, namely integer-valued moving average model (INMA) and Integer-valued Autoregressive Moving Average model (INARMA), and their relationship with cluster point processes, the Cox point process and the dynamic contagion process. We derive the probability generating functionals of INARMA models and compare to that of cluster point processes. The main aim of this paper is to prove that, under a specific parametric setting, INMA and INARMA models are just discrete versions of continuous cluster point processes and hence converge weakly when the length of subintervals goes to zero.

7.1 Introduction

The Hawkes process, which was first introduced in Hawkes (1971a,b), is a self-exciting point process such that its intensity depends on the past of the point process itself. Due to its simplicity and flexibility, the Hawkes process can be viewed as a contagion process and applied in different areas, for example seismology in Ogata (1988), epidemiology in Kim (2011), and sociology in Mohler et al. (2011). It has

gained in popularity in recent years. Finance in particular, is a very popular area to apply Hawkes processes, see Bowsher (2007), Large (2007), Embrechts et al. (2011), Bacry et al. (2012, 2013a,b, 2015), Aït-Sahalia et al. (2015), and Dassios and Zhao (2017a,b). However, in some context such as modelling the credit contagion in Jarrow and Yu (2001), the clustering of defaults is consistent with the Hawkes process, but the default intensity could be impacted exogenously by other factors, which means the distribution of cluster centres may not act as a homogeneous Poisson process in the real financial data. In order to address this, Dassios and Zhao (2011) introduced the dynamic contagion process by generalizing the Hawkes process (with exponential decay kernel) and the Cox process with shot noise intensity (exponential decay kernel) used in Dassios and Jang (2003), which allows the cluster centres act as a stochastic process.

The standard time series models (AR, MA, ARMA, etc.), on the other hand, are used for sequences of real-valued data. A natural question would be whether we can use time series models for count data. An early contribution has been done by Jacobs and Lewis (1978a,b, 1983), who introduced the discrete Autoregressive and Moving average model (DARMA) for stationary discrete time series. However, the correlation structure of DARMA is quite different from the standard time series model. Later, a new model called Integer-valued autoregressive (INAR) time series was defined and examined by McKenzie (1985) and Al-Osh and Alzaid (1987). The idea here is to manipulate the operation between coefficients and variables as well as the innovation terms in a way that the values are always integer. The properties of the INAR model are explored by Al-Osh and Alzaid (1988a), Jin-Guan and Yuan (1991), and McKenzie (1988). The Integer-valued Moving Average model (INMA) was introduced and developed by Al-Osh and Alzaid (1988b), Brännäs and Hall (2001), and Brännäs et al. (2002). They apply the similar idea of the INAR model to a standard MA model.

It seems that no one had studied the connection between point processes and integer-valued time series until Kirchner (2016), who showed that Hawkes point processes are continuous-time versions of Poisson thinning INAR time series with infinite order and vice versa. The author also mentioned that one can introduce the INARMA model by adding the moving average part into the INAR model and hence make a connection to the dynamic contagion process, which is the main motivation of this paper. Basically, we formally define the INMA model in a similar way to Kirchner and prove that the INMA model with infinite order is actually a discrete version of a Cox point process. We then define the INARMA and prove that it is also a discrete version of the dynamic contagion process, as Kirchner expected.

The paper is organized as follows: Section 2 specifies the terminology and reviews the definitions of three cluster point processes, namely the dynamic contagion process, the Cox process and the Hawkes process, and their probability generating functionals. Section 3 reviews the definition of INAR model, defines the INMA model and INARMA model, and derives their probability generating functionals. Section 4 provides further details on the convergence of probability generating functionals between the INARMA models and the cluster point processes. Section 5 establishes the weak convergence result from the INARMA models to their corresponding cluster point processes. Section 6 verifies the convergence theorem by calculating the joint probability generating functions numerically through simulation. A few concluding remarks are in the final section.

7.2 Cluster point processes

In this section, we will first define the space we are working on and provide some terminology and notation concerning the integer-valued random measure. Then, we recall the definitions of three cluster point processes, namely the dynamic contagion process, the Cox process and the Hawkes process. Finally, we derive their probability generating functionals by taking advantage of their cluster representation.

7.2.1 Preliminaries

We will use most of the notation and terminology from Daley and Vere-Jones (2007). Throughout this paper, we work on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{F}$ is the $\sigma-$algebra generating by Ω. A measure μ on the half-line $\mathbb{R}_+$, a complete separable metric space, is boundedly finite if $\mu(A) < \infty$ for every bounded Borel set $A \in \mathcal{B}(\mathbb{R}_+)$. Hence denote $\mathcal{M}^{\#}_{\mathbb{R}_+}$ as the space of all boundedly finite measures and $\mathcal{B}(\mathcal{M}^{\#}_{\mathbb{R}_+})$ as its $\sigma-$algebra.

Definition 7.1. *A point process N on the state space $\mathbb{R}_+$ is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into $(\mathcal{N}^{\#}_{\mathbb{R}_+}, \mathcal{B}(\mathcal{N}^{\#}_{\mathbb{R}_+}))$, $N : \Omega \mapsto \mathcal{N}^{\#}_{\mathbb{R}_+}$, such that $N(A)$ is a integer-valued random variable for each bounded $A \in \mathcal{B}(R_+)$. $\mathcal{N}^{\#}_{\mathbb{R}_+}$ is the family of all boundedly finite integer-valued measure $\mu \in \mathcal{M}^{\#}_{\mathbb{R}_+}$*

For a point process (random measure) $N \in \mathcal{N}^{\#}_{\mathbb{R}_+}$, they are well-defined only on some bounded area. Consequently, the distribution of a point process is completely determined by the finite dimensional distributions, see Proposition 9.2 II in Daley and Vere-Jones (2007)

Definition 7.2. *The finite dimensional distributions of a random measure N are the joint distributions for all finite families of bounded Borel sets $A_1, \ldots, A_k$ of $N(A_1), \ldots, N(A_k)$*

$$F_k(A_1, \ldots, A_k; x_1, \ldots, x_k) = \mathcal{P}\{N(A_i) \leqslant x_i (i = 1, \ldots, k)\}. \tag{7.1}$$

Usually, for a non-negative random measure, one would use the Laplace functional to describe the joint distribution of the random measure. As we work on the space $(\mathcal{N}^{\#}_{\mathbb{R}_+}, \mathcal{B}(\mathcal{N}^{\#}_{\mathbb{R}_+}))$, there are advantages in moving from the Laplace functional to the probability generating functional (p.g.fl)

Definition 7.3. *The probability generating functional (p.g.fl) of a point process N on the complete separable metric space $\mathbb{R}_+$ is defined by*

$$G[h] = \mathbb{E}\left[\exp\left\{\int_{\mathbb{R}_+} \log h(x) N(dx)\right\}\right], \quad h \in \mathcal{V}(\mathbb{R}_+), \tag{7.2}$$

where $\mathcal{V}(\mathbb{R}_+)$ is the class of all real-valued Borel functions h defined on $\mathbb{R}_+$ with $1-h$ vanishing outside some bounded set and satisfying $0 \leqslant h(x) \leqslant 1, \forall x \in \mathbb{R}_+$. Later, we will use $\mathcal{V}_0(\mathbb{R}_+)$, the subset of $\mathcal{V}(\mathbb{R}_+)$ satisfying $\inf_{x \in \mathbb{R}_+} h(x) > 0$

One can always use $G[h]$ to describe F_k by setting $h(x) = h_i, x \in A_i$, where h_i is a constant. Then the $G[h]$ will reduce to the joint probability generating function (joint p.g.f).

$$\begin{aligned} G[h] &= \mathbb{E}\left[\exp\left\{\int_{\mathbb{R}_+} \log h(x) N(dx)\right\}\right] \\ &= \mathbb{E}\left[exp\left(\int_{\cup_{i=1,\ldots,k} A_i} \log h(x) N(dx)\right)\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{k} h_i^{N(A_i)}\right] \end{aligned}$$

In other words, the p.g.fl $G[h]$ is the limit version of the joint p.g.f where the set A_i has the length $dx \to 0$ and $k \to \infty$. When describing the finite dimensional distributions F_k, the p.g.fl and the joint p.g.f are therefore equivalent. For convenience, we will also use the term 'p.g.fl' for those INARMA models to describe their joint p.g.f in section 3.

7.2.2 The dynamic contagion process

We first define a generalized version of the dynamic contagion process as in Dassios and Zhao (2011)

Definition 7.4. *The generalized dynamic contagion process is a cluster point process* $N^{(DCP)}$*, with stochastic intensity* $\lambda^{(DCP)}$ *such that*

$$\lambda_t^{(DCP)} = \sum_{i:c_i<t}^{N_t^*} \Upsilon_i f(t-c_i) + \sum_{i:\tau_i<t}^{N_t^{(DCP)}} \chi_i \eta(t-\tau_i), \tag{7.3}$$

where

- $N_t^* \equiv \{c_i\}_{i=1,2,...}$ *are the arrival times of the Poisson process with the constant rate* $\rho > 0$
- $N_t^{(DCP)} \equiv \{\tau_i\}_{i=1,2,...}$ *are the arrival times of the generalized dynamic contagion process*
- $\{\Upsilon_i\}$ *are i.i.d externally excited jump sizes, realised at times* $\{c_i\}$*, with distribution* $H(x)$*, mean* μ_Υ *and Laplace transform* $\hat{h}(u)$
- $\{\chi_i\}$ *are i.i.d self-exciting jump sizes, realised at times* $\{\tau_i\}$*, with distribution* $G(y)$*, mean* μ_χ *and Laplace transform* $\hat{g}(u)$*. They are independent of* $\{\Upsilon_i\}$
- $f(u)$ *is an Riemann integrable function for any bounded interval in* $\mathbb{R}_+$
- $\eta(u)$ *is another Riemann integrable function for any bounded interval in* $\mathbb{R}_+$

Note that the stationary condition for this point process would be $\int_0^\infty f(u)du < \infty$ and $\mu_\chi \int_0^\infty \eta(u)du < 1$. Following from this definition, we define the other two cluster point processes – the Cox process and the Hawkes process as special cases.

Definition 7.5. *The (Marked) Cox process with shot-noise intensity, also called doubly stochastic process, is a cluster point process* $N^{(C)}$ *with stochastic intensity* $\lambda^{(C)}$ *such that*

$$\lambda_t^{(C)} = \sum_{i:c_i<t}^{N_t^*} \Upsilon_i f(t-c_i). \tag{7.4}$$

It is clear that this is a special case of the dynamic contagion process by letting $\eta(u) = 0, \forall u \in \mathbb{R}_+$. On a bounded area $[0,T]$ where $T > 0$, the process can be considered as a cluster process in which the cluster centres c_i arrive as a homogeneous

Poisson process $N^* \sim Pois(\rho)$. Conditional on the arrival of c_i, we then have a cluster whose size follows $N_t^1 \sim Pois(\Upsilon_i f(t-c_i))$ with $c_i \leqslant t \leqslant T$. Theses clusters are mutually independent and cluster centres are not included in $N^{(C)}$. In order words, the arrivals of cluster centres are indicators that some events will happen around them.

Definition 7.6. *The (Marked) Hawkes process is a self-exciting point process $N^{(H)}$ with stochastic intensity $\lambda^{(H)}$ such that*

$$\lambda_t^{(H)} = \nu + \sum_{i:\tau_i<t}^{N_t^{(H)}} \chi_i \eta(t-\tau_i), \tag{7.5}$$

where ν is a positive constant.

Similarly, this is another special case of the dynamic contagion process by replacing the 'Cox component' in $\lambda_t^{(DCP)}$ by a positive constant ν. From Hawkes and Oakes (1974), the Hawkes process can also be interpreted as a cluster point process. The immigrants (cluster centres) arrive as a homogeneous Poisson process $Pois(\nu)$. Each immigrant generates a Galton–Watson type branching process with expected branching ratio $\mu_\chi \int_0^\infty \eta(u)du$. A cluster is then formed by including all the generations (include the immigrant) from the branching process.

Back to the dynamic contagion process, it is actually a Hawkes process with immigrants arriving as a Cox process rather than a homogeneous Poisson process. Here are the probability generating functionals for these cluster point processes.

Proposition 7.1. *Let $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$ such that $1-z(.)$ vanishes outside $[0,T]$, where $T>0$. The probability generating functional (p.g.fl) of the Cox process $N^{(C)}$ on $[0,T]$ is given by*

$$\begin{aligned} G^{(C)}(z(.)) &= \exp\left\{\rho \int_0^T (F^{(C)}(z(.)|c)-1)dc\right\} \\ F^{(C)}(z(.)|c) &= \hat{h}\left(-\int_0^{T-c} f(u)(z(c+u)-1)du\right). \end{aligned} \tag{7.6}$$

Proof. See appendix 7.A □

Proposition 7.2. *Let $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$ such that $1-z(.)$ vanishes outside $[0,T]$, where $T>0$. The probability generating functional (p.g.fl) of the generalized dynamic*

contagion process $N^{(DCP)}$ *on* $[0,T]$ *is given by*

$$\begin{aligned} G^{(DCP)}(z(.)) &= \exp\left\{\rho\int_0^T\left(\hat{h}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v)-1)f(v)dv\right)-1\right)du\right\} \\ F^{(H)}(z(.)|u) &= z(u)\hat{g}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v)-1)\eta(v)dv\right), \end{aligned} \tag{7.7}$$

where $F^{(H)}(z(.)|u)$ *is the p.g.fl of a cluster generated by an immigrant (cluster centre) arriving at time* u*, and including that immigrant. While* $F(z(.)|u) = F(z_u(.))$ *and* $z_u(.) = z(u+.)$ *is simply the translation of* $z()$*.*

Proof. See appendix 7.B □

Corollary 7.2.1. *Let* $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$ *such that* $1-z(.)$ *vanishes outside* $[0,T]$*, where* $T>0$*. The probability generating functional (p.g.fl) of the Hawkes process* $N^{(H)}$ *on* $[0,T]$ *is given by*

$$\begin{aligned} G^{(H)}(z(.)) &= \exp\left\{\nu\int_0^T\left(F^{(H)}(z(.)|u)-1\right)du\right\} \\ F^{(H)}(z(.)|u) &= z(u)\hat{g}\left(-\int_0^{T-u}(F(z(.)|u+v)-1)\eta(v)dv\right). \end{aligned} \tag{7.8}$$

Proof. This result generally follows from Theorem 2 in Hawkes and Oakes (1974). We can also derive it from Proposition 7.2 by simply letting $\lambda_u^{(C)} = \nu$ in equation 7.34. □

7.3 Poisson thinning Integer-valued time series model

In this section, we will review the Poisson thinning INAR model from Kirchner (2016). Then we will define the INMA and INARMA models in a similar way to the INAR model, and derive their probability generating functionals.

7.3.1 Integer-valued Autoregressive Model - INAR(∞)

We refer to the results of Kirchner (2016) .

Definition 7.7. *The stationary INAR(∞) is defined as*

$$\begin{aligned} X_n &= \sum_{k=1}^{\infty} \alpha_k \circ X_{n-k} + \varepsilon_n \\ &= \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \epsilon_l^{(n,k)} + \varepsilon_n, \end{aligned} \tag{7.9}$$

where

- $\alpha_k \geqslant 0$ *(reproduction coefficients) and the stationary condition* $\sum_{k=1}^{\infty} \alpha_k < 1$
- $\varepsilon_n \overset{i.i.d}{\sim} Pois(\alpha_0)$, *with* $\alpha_0 > 0$ *(immigration parameter)*
- *The thinning operator* $\circ$ *is defined as*

$$\alpha_k \circ X_{n-k} = \sum_{i=1}^{Z_{n-k}} \epsilon_i^{(n,k)} \quad \epsilon_i^{(n,k)} \overset{i.i.d}{\sim} Pois(\alpha_k),$$

  where $\epsilon_l^{(n,k)}$ *are independent over* $n \in \mathbb{Z},\ k \in \mathbb{N},\ i \in \mathbb{N}$

- $\circ$ *operates independently over* $n \in \mathbb{Z},\ k \in \mathbb{N}$

In the early study of the integer-valued time series models, the operator $\circ$ is defined as a binomial thinning operator, which means ϵ_i are Bernoulli random variables. However, Kirchner defines it as a Poisson operator, which will lead to the simpler formulas of probability generating functional. In addition, the p.g.fl derived later can be compared directly to that of the Hawkes process. The following proposition gives the branching representation of the INAR model.

Proposition 7.3. *The INAR(∞) process* X_n *has the following representation*

$$X_n \overset{d}{=} \sum_{i \in \mathbb{Z}} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)}, \tag{7.10}$$

where $F_{n-i}^{(i,j)}$ *are independent over* i, j *and they are the copies of a branching process* F_n *which is defined by*

$$F_n = \sum_{g=0}^{\infty} G_n^{(g)}, \quad n \in \mathbb{Z}. \tag{7.11}$$

The generation G_n are constructed recursively by

$$G_n^{(0)} = 1_{\{n=0\}} \quad G_n^{(g)} = \sum_{k=1}^{n} \alpha_k \circ G_{n-k}^{(g-1)} = \sum_{k=1}^{n} \sum_{m=1}^{G_{n-k}^{(g-1)}} \epsilon_m^{(n,k,g)}, \quad n \in \mathbb{Z},\ g \in \mathbb{N}, \tag{7.12}$$

with $\xi_m^{(n,k,g)}$ are independent over n, k, g, m and also independent of ε_i, $i \in \mathbb{Z}$. Furthermore, we have the following distributional equality for the generic family-process (F_n)

$$(F_n)_{n\in\mathbb{Z}} \stackrel{d}{=} \left(1_{\{n=0\}} + \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} F_{n-1}^{(i,j)} \right). \tag{7.13}$$

Proposition 7.4. *Let $z_\cdot = \{z_i\}_{i=1,\dots,n}$ be a sequence of constants such that $0 < z_i \leqslant 1$. The probability generating functional (p.g.fl) of the INAR sequence $\{X_t\}_{t=1,\dots,n}$ is given by*

$$\begin{aligned} G^{(X_n)}(z_\cdot) &= \exp\left\{ \sum_{i=1}^{n} \alpha_0 (F(z_\cdot|i) - 1) \right\} \\ F^{(X_n)}(z_\cdot|i) &= z_i \exp\left\{ \sum_{k=1}^{n-i} \alpha_k \left(F(z_\cdot|i+k) - 1 \right) \right\}, \end{aligned} \tag{7.14}$$

where $F^{(X_n)}(z_\cdot|t) = F^{(X_n)}(z_{t+\cdot})$ is the p.g.fl of the cluster generated by an immigrant (cluster centre) arriving at time t.

Proof. The (discrete) p.g.fl is given by

$$\begin{aligned} G^{(X_n)}(z_\cdot) = \mathbb{E}[\prod_{t=1}^{n} z_t^{X_t}] &= \mathbb{E} \exp\left\{ \sum_{t=1}^{n} \log z_t \sum_{i=1}^{t} \sum_{j=1}^{\varepsilon_i} F_{t-i}^{(i,j)} \right\} \\ &= \mathbb{E} \exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{\varepsilon_i} \sum_{t=i}^{n} \log z_t F_{t-i}^{(i,j)} \right\}. \end{aligned}$$

The sum $\sum_{t=i}^{n} F_{t-i}^{(i,j)}$ can be interpreted as the cluster, which includes all the generation from time i to time n, generated by one of the immigrants in ε_i. Conditionally on the immigration sequence ε_i and exploiting its independence from the family

process $F_n^{(i,j)}$, we have

$$\begin{aligned} G^{(X_n)}(z_.) &= \prod_{i=1}^{n} \mathbb{E}\left[\prod_{j=1}^{\varepsilon_i} \mathbb{E}\exp\left\{\sum_{t=i}^{n} \log z_t F_{t-i}^{(i,j)}\right\}\right] \\ &= \prod_{i=1}^{n} \mathbb{E}\left[F^{(X_n)}(z_.|i)^{\varepsilon_i}\right] \\ &= \exp\left\{\sum_{i=1}^{n} \alpha_0 (F^{(X_n)}(z_.|i) - 1)\right\}, \end{aligned}$$

where the p.g.fl of the cluster $F^{(X_n)}(z_.|i)$ satisfies the following recursive equation

$$\begin{aligned} F^{(X_n)}(z_.|i) &= \mathbb{E}\exp\left\{\sum_{t=i}^{n} \log z_t F_{t-i}\right\} \\ &= \mathbb{E}\exp\left\{\sum_{t=0}^{n-i} \log z_{i+t}\left(1_{\{t=0\}} + \sum_{k=1}^{t}\sum_{j=1}^{G_k^{(1)}} F_{t-k}^{(k,j)}\right)\right\} \\ &= z_i \prod_{k=1}^{n-i} \mathbb{E}\left[\prod_{j=1}^{G_k^{(1)}} \mathbb{E}\exp\left\{\sum_{t=k}^{n-i} \log z_{i+t} F_{t-k}^{(k,j)}\right\}\right] \\ &= z_i \exp\left\{\sum_{k=1}^{n-i} \alpha_k \left(F^{(X_n)}(z_.|i+k) - 1\right)\right\}. \end{aligned}$$

□

Since the sequence $\{X_t\}_{t=1,\dots,n}$ takes only integer values, if we fix a bounded area $[0,T]$ and let X_t count the number of points for the equal-length area $((t-1)\Delta, t\Delta]$ where $\Delta = \frac{T}{n}$, the p.g.fl of $\{X_t\}_{t=1,\dots,n}$ will look like the discrete version of the p.g.fl of the Hawkes process.

Proposition 7.5. *Consider the following parametric setting.*

- *Fix the bounded area* $[0,T]$, $T < \infty$
- *Choose* $n > 0$, *the number of all subintervals over* $[0,T]$
- *Set the length of subintervals* $\Delta = \frac{T}{n}$, *the immigrant parameter* $\alpha_0 = \nu\Delta$ *and the reproduction coefficient* $\alpha_k = \chi_i \eta(k\Delta)\Delta, k > 0$
- χ_i *are i.i.d random variables corresponding to the cluster centre* $X_{i\Delta}$ *arriving at* $i\Delta$, *with Laplace transform* $\hat{g}(u) = \mathbb{E}[e^{-u\chi_i}]$
- *Let* $z_i = z(i\Delta)$, *where* $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$

Then the probability generating functional of $\{X_t\}_{t=1,\dots,n}$ *becomes*

$$\begin{aligned} G_{(\Delta)}^{(X_n)}(z(.)) &= \exp\left\{\nu\sum_{i=1}^{n}(F(z(.)|i\Delta)-1)\Delta\right\} \\ F^{(X_n)}(z(.)|i\Delta) &= z(i\Delta)\hat{g}\left(-\sum_{k=1}^{n-i}(F(z(.)|(i+k)\Delta)-1)\eta(k\Delta)\Delta\right). \end{aligned} \tag{7.15}$$

Proof. By substituting $\alpha_k = \chi_i\eta(k\Delta)\Delta, k > 0$ into Proposition 7.4, the p.g.fl of the cluster $F^{(X_n)}(z.|i) = F^{(X_n)}(z(.)|i\Delta)$ becomes

$$\begin{aligned} F^{(X_n)}(z(.)|i\Delta) &= z(i\Delta)\prod_{k=1}^{n-i}\mathbb{E}\left[\prod_{j=1}^{G_k^{(1)}}\mathbb{E}\exp\left\{\sum_{t=k}^{n-i}\log z((i+t)\Delta)F_{t-k}^{(k,j)}\right\}\right] \\ &= z(i\Delta)\mathbb{E}\left[\exp\left\{\sum_{k=1}^{n-i}\chi_i\eta(k\Delta)\Delta\left(F^{(X_n)}(z(.)|(i+k)\Delta)-1\right)\right\}\right] \\ &= z(i\Delta)\hat{g}\left(-\sum_{k=1}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta)-1)\eta(k\Delta)\Delta\right). \end{aligned}$$

By substituting $\alpha_0 = \nu\Delta$, the whole p.g.fl of the INAR sequence $\{X_t\}_{t=1,\dots,n}$ becomes

$$\begin{aligned} G_{(\Delta)}^{(X_n)}(z(.)) &= \exp\left\{\sum_{i=1}^{n}\alpha_0(F^{(X_n)}(z_.|i)-1)\right\} \\ &= \exp\left\{\nu\sum_{i=1}^{n}(F^{(X_n)}(z(.)|i\Delta)-1)\Delta\right\}. \end{aligned}$$

□

7.3.2 Integer-valued Moving Average model

Definition 7.8. *The stationary Poisson thinning INMA(*∞*) model is defined as*

$$\begin{aligned} Y_n &= \sum_{k=0}^{\infty}\beta_k\circ\xi_{n-k} \\ &= \beta_0\circ\xi_n + \beta_1\circ\xi_{n-1} + \cdots + \beta_{n-1}\circ\xi_1, \end{aligned} \tag{7.16}$$

where

- $\beta_k \geqslant 0$ *are some non-negative coefficients,* $\sum_{k=0}^{\infty}\beta_k < \infty$ *and* $\lim_{k\to\infty}\beta_k = 0$

- ξ_k *are i.i.d and follow* $Pois(\mu)$ *with* $\mu > 0$
- *The thinning operator* $\circ$ *is defined as*

$$\beta_k \circ \xi_{n-k} = \sum_{i=1}^{\xi_{n-k}} u_i^{(n,k)}, \quad u_i^{(n,k)} \overset{i.i.d}{\sim} Pois(\beta_k),$$

where $u_i^{(n,k)}$ *are independent over* $n \in \mathbb{N},\ k \in \mathbb{N},\ i \in \mathbb{N}$.

The parameters β_k and μ have a similar interpretation to those in the INAR model. β_k are reproduction coefficients while μ is the arrival intensity of cluster centre rather than 'immigrants' because it is not counted by the system Y_n. From this model point of view, we can regard ξ_n as the cluster centres. They enter the system starting at time n and trigger other events at each time period ($n \to \sum_{i=1}^{\xi_n} u_i^{(n,0)}, n+1 \to \sum_{i=1}^{\xi_n} u_i^{(n+1,1)}, \dots$). Y_n is then a counting variable to report the total number of the triggered events from $\xi_n, \xi_{n-1}, \dots, \xi_1$ over the current time period n. Here are two assumptions we need before proceeding to its probability generating functional.

Assumption 7.3.1. *The thinning operations* $\beta_k \circ \xi_{n-k}$ *are mutually independent for* $n \in \mathbb{N},\ k \in \mathbb{N}$.

Assumption 7.3.2. $u_i^{(n,k)}$ *are mutually independent of each other for* $n \in \mathbb{N},\ k \in \mathbb{N}$.

The second assumption means that the number of events $u_i^{(t,k)}$, triggered by one of the cluster centre in ξ_{t-k} and counted by the system Y_t, will not affect the number of events $u_i^{(t+j,k+j)}$, triggered by the same cluster centre and counted by the system Y_{t+j} of any future time $j > 0$.

Proposition 7.6. *Let* $z_\cdot = \{z_i\}_{i=1,\dots,n}$ *be a sequence of constants such that* $0 < z_i \leqslant 1$. *The probability generating functional (p.g.fl) of the INMA sequence* $\{Y_t\}_{t=1,\dots,n}$ *is given by*

$$\begin{aligned} G^{(Y_n)}(z_\cdot) &= \exp\left\{\mu \sum_{t=1}^{n} (F^{(Y_n)}(z_\cdot|t) - 1)\right\} \\ F^{(Y_n)}(z_\cdot|t) &= \exp\left\{\sum_{k=1}^{n-t+1} \beta_{k-1}(z_{t+k-1} - 1)\right\}. \end{aligned} \tag{7.17}$$

Proof. The aggregated process $S_n = \sum_{t=1}^{n} Y_t$ is actually a cluster point process such

that

$$\begin{aligned} S_n &= \sum_{t=1}^{n} Y_t = \sum_{t=1}^{n}\sum_{k=0}^{t-1} \beta_k \circ \xi_{t-k} \\ &= \sum_{t=1}^{n}\sum_{k=0}^{n-t} \beta_k \circ \xi_t \\ &= \sum_{t=1}^{n}\sum_{i=1}^{\xi_t} (u_i^{(t,0)} + u_i^{(t+1,1)} + \cdots + u_i^{(n,n-t)}), \quad u_i^{(t,k)} \sim Pois(\beta_k) \\ &\stackrel{d}{=} \sum_{t=1}^{n}\sum_{i=1}^{\xi_t} u_i^t, \quad u_i^t \sim Pois(\sum_{k=0}^{n-t} \beta_k). \end{aligned} \tag{7.18}$$

The last equality follows from the independence of the Poisson random variables. It is now clear that the aggregated process S_n is a cluster process such that

- ξ_t generates the cluster centres independently.
- u_i^t is a cluster generated by one of the cluster centre from ξ_t, with the size of cluster (exclude the cluster centre) following $Pois(\sum_{k=0}^{n-t} \beta_k)$

The (discrete) p.g.fl of Y_t is defined as

$$G(z_\cdot) = \mathbb{E}[\prod_{j=0}^{n} z_j^{Y_j}].$$

Now we can derive the p.g.fl of this process by following the similar argument in Proposition 7.1. Conditionally on the arrivals of cluster centres generated by ξ_t, the p.g.fl of cluster u_i^t is

$$F^{(Y_n)}(z_\cdot|t) = \mathbb{E}[\prod_{k=0}^{n-t} z_{t+k}^{u_i^{(t+k,k)}}] = \exp\left\{\sum_{k=0}^{n-t} \beta_k (z_{t+k} - 1)\right\}.$$

The cluster centres generated by ξ_t are mutually independent. Then the p.g.fl of $\sum_{i=1}^{\xi_t} u_i^t$ is

$$G_t(z_\cdot) = \mathbb{E}[F^{(Y_n)}(z_\cdot|t)^{\xi_t}] = \exp\left\{\mu(F^{(Y_n)}(z_\cdot|t) - 1)\right\}.$$

Clusters centres generated by $\{\xi_t\}_{t=1,\ldots,n}$ are also mutually independent. Finally the p.g.fl of Y_t is

$$G^{(Y_n)}(z_\cdot) = \prod_{t=1}^{n} G_t(z) = \exp\left\{\mu \sum_{t=1}^{n} (F^{(Y_n)}(z_\cdot|t) - 1)\right\}.$$

□

Similar to the INAR model, due to the integer-valued nature of the INMA model, if we fix an bounded area $[0,T]$ and let Y_t counts the number of points for the equal-length area $((t-1)\Delta, t\Delta]$ with $\Delta = \frac{T}{n}$, the p.g.fl of the INMA sequence $\{Y_t\}_{t=1,\dots,n}$ will look like the discrete version of the p.g.fl of the Cox process under some specific parametric setting.

Proposition 7.7. *Consider the following parametric setting.*

- *Fix the bounded area* $[0,T]$, $T < \infty$
- *Choose* $n > 0$, *the number of all subintervals over* $[0,T]$
- *Set the length of subintervals* $\Delta = \frac{T}{n}$, $\mu = \rho\Delta$ *and* $\beta_k = \Upsilon_t f(k\Delta)\Delta, k \geqslant 0$
- Υ_t *are i.i.d random variables corresponding to the cluster centre* $\xi_{t\Delta}$ *arriving at* $t\Delta$, *with the Laplace transform* $\hat{h}(u) = \mathbb{E}[e^{-u\Upsilon_t}]$
- *Let* $z_k = z(k\Delta)$ *where* $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$

Then the probability generating functional of the sequence $\{Y_t\}_{t=1,\dots,n}$ *becomes*

$$\begin{aligned} G_{(\Delta)}^{(Y_n)}(z(.)) &= \exp\left\{\rho\sum_{t=1}^{n}(F^{(Y_n)}(z(.)|t\Delta) - 1)\Delta\right\} \\ F^{(Y_n)}(z(.)|t\Delta) &= \hat{h}\left(-\sum_{k=1}^{n-t+1} f(k\Delta)(z((t+k-1)\Delta) - 1)\Delta\right). \end{aligned} \tag{7.19}$$

Proof. By substituting $\beta_k = \Upsilon_t f(k\Delta)\Delta, k \geqslant 0$ into Proposition 7.6, the p.g.fl of the cluster part $F^{(Y_n)}(z_.|t) = F^{(Y_n)}(z(.)|t\Delta)$ becomes

$$\begin{aligned} F^{(Y_n)}(z(.)|t\Delta) &= \mathbb{E}[\prod_{k=0}^{n-t} z_{t+k}^{u_i^{(t+k,k)}}] \\ &= \mathbb{E}\left[\exp\left\{\sum_{k=0}^{n-t}\Upsilon_t f(k\Delta)\Delta(z_{t+k}-1)\right\}\right] \\ &= \hat{h}\left(-\sum_{k=1}^{n-t+1} f(k\Delta t)(z((t+k-1)\Delta) - 1)\Delta\right). \end{aligned}$$

Then substituting $\mu = \rho\Delta$, the p.g.fl of the INMA sequence $\{Y_t\}_{t=1,\ldots,n}$ becomes

$$\begin{aligned} G_{(\Delta)}^{(Y_n)}(z(.)) &= \exp\left\{\mu \sum_{t=1}^{n} (F^{(Y_n)}(z(.)|t\Delta) - 1)\right\} \\ &= \exp\left\{\rho \sum_{t=1}^{n} (F^{(Y_n)}(z(.)|t\Delta) - 1)\Delta\right\}. \end{aligned}$$

□

7.3.3 Integer-valued Autoregressive Moving Average model

Definition 7.9. *The stationary Poisson thinning INARMA(∞,∞) model is defined as*

$$\begin{aligned} Z_n &= \sum_{k=1}^{\infty} \alpha_k \circ Z_{n-k} + Y_n \\ &= \sum_{k=1}^{\infty} \alpha_k \circ Z_{n-k} + \sum_{j=0}^{\infty} \beta_j \circ \xi_{n-j} \\ &= \alpha_1 \circ Z_{n-1} + \cdots + \alpha_{n-1} \circ Z_1 + \beta_0 \circ \xi_t + \cdots + \beta_{n-1} \circ \xi_1, \end{aligned} \tag{7.20}$$

where

- $\xi_t \overset{i.i.d}{\sim} Pois(\mu)$.
- α_i *and* β_i *are positive coefficients, and* $\sum_{i=1}^{\infty} \alpha_i < 1$. $\sum_{i=0}^{\infty} \beta_i < \infty$, $\lim_{i\to\infty} \beta_i = 0$.
- *The thinning operator* $\circ$ *is defined as*

$$\alpha_k \circ Z_{n-k} = \sum_{i=1}^{Z_{n-k}} \epsilon_i^{(n,k)} \quad \epsilon_i^{(n,k)} \overset{i.i.d}{\sim} Pois(\alpha_k)$$

$$\beta_k \circ \xi_{n-k} = \sum_{i=1}^{\xi_{n-k}} u_i^{(n,k)}, \quad u_i^{(n,k)} \overset{i.i.d}{\sim} Pois(\beta_k).$$

- $\epsilon_i^{(n,k)}$ *and* $u_i^{(n,k)}$ *are mutually independent over* $n \in \mathbb{N}, i \in \mathbb{N},\ k \in \mathbb{N}$.

The INARMA model simply combines the INAR components and the INMA components from previous sections. It is a generalized INAR model whose immigrants process ε_n *is replaced by the INMA model* Y_n.

Proposition 7.8. *Let $z_\cdot = \{z_i\}_{i=1,\dots,n}$ be a sequence of constants such that $0 < z_i \leqslant 1$. The probability generating functional (p.g.fl) of the INARMA sequence $\{Z_t\}_{t=1,\dots,n}$ is given by*

$$\begin{aligned} G^{(Z_n)}(z_\cdot) &= \exp\left\{\mu \sum_{i=1}^{n}\left(\exp\left\{\sum_{k=0}^{n-i}\beta_k(F^{(X_n)}(z_\cdot|i+k)-1)\right\}-1\right)\right\} \\ F^{(X_n)}(z_\cdot|i) &= z_i \exp\left\{\sum_{k=1}^{n-i}\alpha_k\left(F^{(X_n)}(z_\cdot|i+k)-1\right)\right\}, \end{aligned} \tag{7.21}$$

where $F^{(X_n)}(z_\cdot|t)$ is the p.g.fl of the cluster generated by an immigrant (cluster centre) arriving at time t, and including that immigrant. While $F^{(X_n)}(z_\cdot|t) = F^{(X_n)}(z_{t+\cdot})$ is simply the translation of $z_\cdot$.

Proof. The $F^{(X_n)}(z_\cdot|i)$ is exactly the same as the one in INAR model, because this is the cluster generated by the autoregressive structure in the INARMA model and it is irrelevant to Y_i. Hence we can apply the result directly from Proposition 7.4

$$F^{(X_n)}(z_\cdot|i) = z_i \exp\left\{\sum_{k=1}^{n-i}\alpha_k\left(F^{(X_n)}(z_\cdot|i+k)-1\right)\right\}.$$

Then we can apply a similar argument to the INAR model such that

$$\begin{aligned} G^{(Z_n)}(z_\cdot) &= \prod_{i=1}^{n}\mathbb{E}\left[\prod_{j=1}^{Y_i}\mathbb{E}\exp\left\{\sum_{t=i}^{n}\log z_t F_{t-i}^{(i,j)}\right\}\right] \\ &= \prod_{i=1}^{n}\mathbb{E}\left[F^{(X_n)}(z_\cdot|i)^{Y_i}\right]. \end{aligned}$$

Now apply the p.g.fl of the INMA model from Proposition 7.6

$$\begin{aligned} G^{(Z_n)}(z_\cdot) &= \prod_{i=1}^{n}\mathbb{E}\left[F^{(X_n)}(z_\cdot|i)^{Y_i}\right] \\ &= \exp\left\{\mu\sum_{i=1}^{n}\left(\exp\left\{\sum_{k=0}^{n-i}\beta_k(F^{(X_n)}(z_\cdot|i+k)-1)\right\}-1\right)\right\}. \end{aligned}$$

□

Similar to the INAR model, due to the integer-valued nature of INARMA model, if we fix a bounded area $[0,T]$ and let Z_t counts the number of points for the equal-length area $((t-1)\Delta, t\Delta]$ with $\Delta = \frac{T}{n}$, the p.g.fl of the INARMA sequence

$\{Z_t\}_{t=1,\dots,n}$ will look like the discrete version of the p.g.fl of the generalized dynamic contagion process.

Theorem 7.1. *Consider the following parametric setting.*

- *Fixed the terminal time T, $T < \infty$*
- *Choose $n > 0$, the number of all subintervals over $[0, T]$*
- *Set the length of subintervals $\Delta = \frac{T}{n}$, the parameters of INAR part $\alpha_k = \chi_i \eta(k\Delta)\Delta, k > 0$ and the parameters of INMA part $\mu = \rho\Delta t$, $\beta_j = \Upsilon_i f(j\Delta)\Delta, j \geqslant 0$*
- *Υ_i are i.i.d random variables corresponding to each cluster centre $\xi_{i\Delta}$ arriving at $i\Delta$, with the Laplace transform $\hat{h}(u) = \mathbb{E}[e^{-u\Upsilon_i}]$*
- *χ_i are i.i.d random variables corresponding to each INAR cluster centre $Z_{i\Delta}$ arriving at $i\Delta$, with the Laplace transform $\hat{g}(u) = \mathbb{E}[e^{-u\chi_i}]$*
- *Let $z_t = z(t\Delta)$ with $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$.*

The probability generating functional of the INARMA sequence $\{Z_t\}_{t=1,\dots,n}$ becomes

$$G_{(\Delta)}^{(Z_n)}(z(.)) = \exp\left\{\rho \sum_{i=1}^{n} \left(\hat{h}\left(-\sum_{k=0}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta) - 1)f(k\Delta)\Delta\right) - 1\right)\Delta\right\}$$

$$F^{(X_n)}(z(.)|i\Delta) = z(i\Delta)\hat{g}\left(-\sum_{k=1}^{n-i}\left(F^{(X_n)}(z(.)|(i+k)\Delta) - 1\right)\eta(k\Delta)\Delta\right). \tag{7.22}$$

Proof. By substituting $\alpha_k = \chi_i \eta(k\Delta)\Delta$ into Proposition 7.8, the p.g.fl $F^{(X_n)}(z_.|i) = F(z(.)|i\Delta)$ is exactly the same as the one in Proposition 7.5

$$\begin{aligned} F^{(X_n)}(z(.)|i\Delta) &= z(i\Delta)\mathbb{E}\left[\exp\left\{\sum_{k=1}^{n-i}\chi_i\eta(k\Delta)\Delta\left(F^{(X_n)}(z(.)|(i+k)\Delta) - 1\right)\right\}\right] \\ &= z(i\Delta)\hat{g}\left(-\sum_{k=1}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta) - 1)\eta(k\Delta)\Delta\right). \end{aligned}$$

By substituting $\beta_k = \Upsilon_i f(k\Delta)\Delta$, the p.g.fl of the whole INARMA sequence $\{Z_t\}_{t=1,\dots,n}$ becomes the p.g.fl of INMA sequence $\{Y_t\}_{t=1,\dots,n}$. Then we can apply Proposition

7.7

$$
\begin{aligned}
G_{(\Delta)}^{(Z_n)}(z(.)) &= G_{(\Delta)}^{(Y_n)}(F^{(X_n)}(z(.)|i\Delta)) \\
&= \exp\left\{\rho\sum_{i=1}^{n}\left(\hat{h}\left(-\sum_{k=0}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta)-1)f(k\Delta)\Delta\right)-1\right)\Delta\right\},
\end{aligned}
$$

where the $z(.)$ is replaced by $F^{(X_n)}(z(.)|i\Delta)$ in $G_{(\Delta)}^{(Y_n)}$. □

7.4 Convergence of probability generating functionals

In this section, we will prove the convergence results of the p.g.fl.s between the INARMA models and the cluster point processes.

7.4.1 Dynamic contagion process and INARMA model

The p.g.fl of the generalized dynamic contagion process is given by

$$
\begin{aligned}
G^{(DCP)}(z(.)) &= \exp\left\{\rho\int_0^T\left(\hat{h}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v)-1)f(v)dv\right)-1\right)du\right\} \\
F^{(H)}(z(.)|u) &= z(u)\hat{g}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v)-1)\eta(v)dv\right).
\end{aligned}
\tag{7.23}
$$

The p.g.fl of the INARMA model with specific parametric setting in theorem 7.1 is given by

$$
\begin{aligned}
G_{(\Delta)}^{(Z_n)}(z(.)) &= \exp\left\{\rho\sum_{i=1}^{n}\left(\hat{h}\left(-\sum_{k=0}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta)-1)f(k\Delta)\Delta\right)-1\right)\Delta\right\} \\
F^{(X_n)}(z(.)|i\Delta) &= z(i\Delta)\hat{g}\left(-\sum_{k=1}^{n-i}(F^{(X_n)}(z(.)|(i+k)\Delta)-1)\eta(k\Delta)\Delta\right).
\end{aligned}
\tag{7.24}
$$

Lemma 7.1. *If $r(u)$ is an Riemann integrable function over an interval $[a,b]$ such that $r(u)$ is bounded and the set D, the discontinuities of $r(u)$, has Lebesgue measure*

0, then there exist positive constants M and k satisfy the following inequality

$$\left|\int_a^b r(u)du - R_n\right| \leqslant M\Delta^k \sim O(\Delta^k), \quad n > 0, \tag{7.25}$$

where

- *n - number of subintervals over* $[a, b]$ *which has the partition* $\{x_0, x_1, \ldots, x_n\}$ *such that* $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$
- $R_n = \sum_{i=1}^n r(t_i)\Delta_i$, *where* $x_i \in [x_{i-1}, x_i]$, $\Delta_i = x_i - x_{i-1}$ *and* $\Delta = \max_{i=1,\ldots,n} \Delta_i$

Proof. From the definition of Riemann integral, for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left|\int_a^b r(u)du - R_n\right| < \epsilon, \quad for\ \Delta < \delta.$$

Then conversely, for every choice of δ, there exists ϵ such that the above inequality holds and it converges to 0 when $\delta \to 0$ from which we can infer that the ϵ is the function of Δ with a positive power. Then we let $\delta = \Delta$ and let $\epsilon = M\Delta^k > 0$ for some positive M and k such that the above inequality also holds for the case of equality . □

Proposition 7.9. *Let* Θ *be the parameter space to specify the generalized dynamic contagion process and the INARMA model and* $z() \in \mathcal{V}_0(\mathbb{R}_+)$. *There exist a positive constant k such that the rate of convergence for the absolute difference of the log p.g.fl.s between the generalized dynamic contagion process and the INARMA model is given by*

$$\begin{aligned} &D^{(DCP)}(z(.), \Delta|\Theta) = \left|\log G^{(DCP)}(z(.)) - \log G^{(Z_n)}_{(\Delta)}(z(.))\right| \sim O(\Delta^k) \\ &\lim_{n\to\infty} D^{(DCP)}(z(.), \Delta|\Theta) = 0. \end{aligned} \tag{7.26}$$

Proof. See the appendix 7.C. □

Corollary 7.4.1. *Let* Θ *be the parameter space to specify the Cox process and the INMA model and* $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$. *There exist a positive constant k such that the rate of convergence for the absolute difference of the log p.g.fl.s between the Cox process and the INMA model is given by*

$$\begin{aligned} &D^{(C)}(z(.), \Delta|\Theta) = \left|\log G^{(C)}(z(.)) - \log G^{(Y_n)}_{(\Delta)}(z(.))\right| \sim O(\Delta^k) \\ &\lim_{n\to\infty} D^{(C)}(z(.), \Delta|\Theta) = 0. \end{aligned} \tag{7.27}$$

Proof. See appendix 7.D. □

Corollary 7.4.2. *Let Θ be the parameter space to specify the Hawkes process and the INAR model and $z() \in \mathcal{V}_0(\mathbb{R}_+)$. There exist a positive constant k such that the rate of convergence for the absolute difference of the log p.g.fl.s between the Hawkes process and the INAR model is given by*

$$\begin{aligned} &D^{(H)}(z(.), \Delta|\Theta) = \left|\log G^{(H)}(z(.)) - \log G^{(X_n)}_{(\Delta)}(z(.))\right| \sim O(\Delta^k) \\ &\lim_{n\to\infty} D^{(H)}(z(.), \Delta|\Theta) = 0. \end{aligned} \tag{7.28}$$

Proof. See appendix 7.E. □

7.5 Links between the INARMA models and the cluster point processes

In this section, we will construct a family of random measures $\{N_n\}_{n=1,2,...}$ on $\mathcal{B}(\mathcal{N}^{\#}_{\mathbb{R}_+})$ by aggregating the integer-valued time series and explain how the discrete time models can mimic the behaviour of those continuous time cluster point processes N. We prove that, under the weak convergence theorem, N_n will converge weakly to N as $n \to \infty$.

7.5.1 Preliminaries and definition

As discussed in the previous section, we can always fix a bounded area $[0, T]$ and choose a number $n > 0$, large enough. Then a continuous point process $N((0, T])$ can be treated as the sum of the bin-size count $\{N((t-1)\Delta, t\Delta])\}_{t=1,...,n}$ with $\Delta = \frac{T}{n}$. Conversely, for example the INAR model, we let the sequence $\{X_t\}_{t=1,...,n}$ be the measures for the bin-size count $\{N((t-1)\Delta, t\Delta])\}_{t=1,...,n}$. Hence if we specify the parameters in integer-valued time series models carefully and if n is large enough, we would expect the aggregation of the integer-valued time series can approximate the continuous cluster point process.

Definition 7.10. *For $n > 0$, let $\{X_t\}_{t=1,...,n}$,$\{Y_t\}_{t=1,...,n}$ and $\{Z_t\}_{t=1,...,n}$ be the INAR sequence, the INMA sequence and the INARMA sequence defined in section 3 with*

the parametric setting $\Delta = \frac{T}{n}$, $\alpha_0 = \nu\Delta, \alpha_k = \chi_i\eta(k\Delta)\Delta$ *for* $k > 0$, $\beta_j = \Upsilon_i f(j\Delta)\Delta$ *for* $j \geqslant 0$ *and* $\mu = \rho\Delta$. *Define the following three families of point processes,*

$$\begin{aligned} N_n^{(H)}(A) &= \sum_{t:t\Delta\in A} X_t \\ N_n^{(C)}(A) &= \sum_{t:t\Delta\in A} Y_t \\ N_n^{(DCP)}(A) &= \sum_{t:t\Delta\in A} Z_t \end{aligned} \tag{7.29}$$

where A *is a bounded set in* $\mathcal{B}(\mathbb{R}_+)$ *and* T *is a constant such that* $T \geqslant \sup A$. *The joint distribution of these point processes are uniquely determined by their p.g.fl.s derived in the section 3.*

The idea here is basically followed from Kirchner (2016). To prove the weak convergence, he defined the INAR model and construct a family of point processes $N^{(\Delta)}$ by aggregating the INAR sequence over $A \in \mathcal{B}(\mathbb{R})$, the Borel σ-algebra on $\mathbb{R}$. Then he proved the weak convergence of $N^{(\Delta)}$ to the Hawkes process N from the definition point of view, see definition 5 and Theorem 2 in Kirchner (2016). He also mentioned this can be proved in a different way by showing the convergence of the Laplace functional of $N^{(\Delta)}$. In our case, we will use probability generating functionals.

7.5.2 Weak convergence

From definition 7.1 in section 2 and Proposition 9.2.II in Daley and Vere-Jones (2007), we can say that the distribution of a random measure (point process) ζ on $(\mathcal{N}_{\mathbb{R}_+}^{\#}, \mathcal{B}(\mathcal{N}_{\mathbb{R}_+}^{\#}))$ is completely determined by its finite-dimensional distributions. Then for the weak convergence of random measure on $\mathcal{N}_{\mathbb{R}_+}^{\#}$, it is sufficient to prove the convergence of finite dimensional distributions, which is established by Theorem 11.1.VII in Daley and Vere-Jones (2007).

Proposition 7.10. *Let* $\mathcal{X}$ *be a complete separable metric space and let* $\mathcal{P}$, $\{\mathcal{P}_n\}$ *be distributions on* $(\mathcal{M}_{\mathcal{X}}^{\#}, \mathcal{B}(\mathcal{M}_{\mathcal{X}}^{\#}))$. *Then* $\mathcal{P}_n \to \mathcal{P}$ *weakly if and only if the finite-dimensional distributions of* $\mathcal{P}_n$ *converge weakly to those of* $\mathcal{P}$.

In our case, the state space is $\mathcal{X} = \mathbb{R}_+$. Also, there is one-to-one mapping from finite dimensional distributions to its probability generating functional. Hence it is sufficient to prove the convergence of the p.g.fl.s between point processes. This is

confirmed by another Proposition 11.1.VIII in Daley and Vere-Jones (2007). We only write down part of it here.

Proposition 7.11. *Each of the following conditions is equivalent to the weak convergence $\mathcal{P}_n \to \mathcal{P}$, assuming the function f ranges over the space of continuous functions vanishing outside a bounded set.*

- *The distribution of $\int_{\mathcal{X}} f d\zeta$ under $\mathcal{P}_n$ converges weakly to its distribution under $\mathcal{P}$*
- *For point process, the p.g.fl.s $G_n[z]$ converge to $G[z]$ for each continuous $z \in \mathcal{V}_0(\mathcal{X})$*

Before establishing the convergence theorem, we need to first show the probability measures of those point processes defined in 7.10 are uniformly tight. Here we refer and combine the results of Lemma 1 and 2 in Kirchner (2016). We also derive a similar one for $N_n^{(C)}$ and $N_n^{(DCP)}$.

Lemma 7.2. *For any bounded interval $[a, b]$ on $\mathbb{R}_+$, we can always find a constant $T > b$ and define $\Delta = \frac{T}{n} \in (0, \delta)$ for some constant $\delta > 0$ as long as $n > \lceil \frac{T}{\delta} \rceil$. Let $N_n^{(H)}$ be the point process defined in 7.10 Then there exists a constant $B^{(H)}$ such that*

$$\begin{aligned} \mathbb{E}[N_n^{(H)}([a,b])] &< (b-a+2\delta)\nu B^{(H)} \\ B^{(H)} &= \begin{cases} (1-K)^{-1}, & \text{if } K < 1 \\ (1+K+K^2+\cdots+K^m), & \text{otherwise} \end{cases} \\ K &= \mu_\chi \sum_{k=1}^{\infty} \eta(k\Delta)\Delta \end{aligned} \tag{7.30}$$

Proof. The coefficients $(b-a+2\delta)\nu$ denote the upper bound of the expected number of immigrants over the fixed time interval $[a, b]$, whose derivation is given in Kirchner (2016). In the stationary case where the branching ratio $K < 1$. The expected size of a cluster for INAR(∞) over a long time horizon is evaluated as $(1+K+K^2+\dots) = (1-K)^{-1}$. In the non-stationary case, since the offspring is produced by Poisson distribution, there is a positive waiting time before a new generation is produced. So over the bounded interval $[a, b]$, there exists a constant $m > 0$ and the size of a cluster is the sum of m generations $(1 + K + K^2 + \cdots + K^m)$ □

Lemma 7.3. *For any bounded interval $[a, b]$ on $\mathbb{R}_+$, we can always find a constant $T > b$ and define $\Delta = \frac{T}{n} \in (0, \delta)$ for some constant $\delta > 0$ as long as $n > \lceil \frac{T}{\delta} \rceil$. Let*

$N_n^{(C)}$ and $N_n^{(DCP))}$ be the point processes defined in 7.10. Then there exist constants $B^{(H)}$ and $L(T)$ such that

$$\begin{aligned}\mathbb{E}[N_n^{(C)}([a,b])] &< (b-a+2\delta)\rho L(T)\\ \mathbb{E}[N_n^{(DCP)}([a,b])] &< (b-a+2\delta)\rho L(T)B^{(H)},\end{aligned} \tag{7.31}$$

where $L(T) = \mu_\Upsilon(\int_0^T f(t)dt+c)$. The constant c is defined as $c = \left|\int_0^T f(t)dt - \sum_{k=0}^{n-1} f(k\Delta)\Delta\right|$

Proof. From the definition of INMA model, the expectation is

$$\begin{aligned}\mathbb{E}[Y_t] &= \mathbb{E}[\beta_0 \circ \xi_t + \cdots + \beta_{t-1} \circ \xi_1]\\ &= \mathbb{E}[\xi_i]\sum_{k=0}^{n-1}\mathbb{E}[\beta_k]\\ &= \rho\Delta\sum_{k=0}^{n-1}\mu_\Upsilon f(k\Delta)\Delta\\ &\leqslant \rho\Delta\mu_\Upsilon\left(\int_0^T f(t)dt + c\right)\\ &\leqslant \rho L(T)\Delta.\end{aligned}$$

The number of subintervals over $[a,b]$ is $[\frac{b-a}{\Delta}]+1 < \frac{b-a}{\Delta}+2$. Finally we have

$$\begin{aligned}\mathbb{E}[N_n^{(C)}([a,b])] &= \sum_{t,t\Delta\in[a,b]}\mathbb{E}[Y_t]\\ &\leqslant \left(\left[\frac{b-a}{\Delta}\right]+1\right)\rho L(T)\Delta\\ &< \left(\frac{b-a}{\Delta}+2\right)\rho L(T)\Delta\\ &< (b-a+2\delta)\rho L(T).\end{aligned}$$

The upper bound for $\mathbb{E}[N_n^{(DCP)}([a,b])]$ can be derived similarly as that of $\mathbb{E}[N_n^{(H)}([a,b])]$. We need to replace ν by $\rho L(T)$ □

Lemma 7.4. *The families of the probability measures $\mathcal{P}_n^{(C)}, \mathcal{P}_n^{(H)}, \mathcal{P}_n^{(DCP)}$ on $\left(\mathcal{N}_{\mathbb{R}_+}^{\#}, \mathcal{B}(\mathcal{N}_{\mathbb{R}_+}^{\#})\right)$ corresponding to the point processes $N_n^{(C)}, N_n^{(H)}, N_n^{(DCP)}$ respectively are uniformly tight.*

Proof. For any bounded interval $[a,b]$ on $\mathbb{R}_+$, we can always find a constant $T > b$ and define $\Delta = \frac{T}{n}$, such that $\Delta \in (0,\delta)$ for some constant $\delta > 0$ as long as $n > \left[\frac{T}{\delta}\right]$.

To show the tightness, for every $\epsilon > 0$, we can let $M_\epsilon^{(H)} = (b-a+2\delta)\frac{\nu B^{(H)}}{\epsilon}$, $M_\epsilon^{(C)} = (b-a+2\delta)\rho L(T)\frac{1}{\epsilon}$ and $M_\epsilon^{(DCP)} = (b-a+2\delta)\frac{\rho L(T) B^{(H)}}{\epsilon}$ such that

$$
\begin{aligned}
P(N_n^{(H)}([a,b]) > M_\epsilon^{(H)}) &\leqslant \frac{\mathbb{E}[N_n^{(H)}([a,b])]}{M_\epsilon^{(H)}} < (b-a+2\delta)\frac{\nu B^{(H)}}{M_\epsilon^{(H)}} = \epsilon \\
P(N_n^{(C)}([a,b]) > M_\epsilon^{(C)}) &\leqslant \frac{\mathbb{E}[N_n^{(C)}([a,b])]}{M_\epsilon^{(H)}} < \frac{(b-a+2\delta)\rho L(T)}{M_\epsilon^{(H)}} = \epsilon \\
P(N_n^{(DCP)}([a,b]) > M_\epsilon^{(DCP)}) &\leqslant \frac{\mathbb{E}[N_n^{(DCP)}([a,b])]}{M_\epsilon^{(H)}} < (b-a+2\delta)\frac{\rho L(T) B^{(H)}}{M_\epsilon^{(DCP)}} = \epsilon
\end{aligned}
$$

Here we apply the Markov inequality. □

Theorem 7.2. *Let $N^{(H)}, N^{(C)}, N^{(DCP)}$ be the Hawkes process, the Cox process and the generalized dynamic contagion process defined in section 2. For $n > 0$, let $N_n^{(H)}$, $N_n^{(C)}$ and $N_n^{(DCP)}$ be the point processes defined in 7.10. Then we have the following weak convergence results*

$$
\begin{aligned}
&N_n^{(H)} \xrightarrow{w} N^{(H)} \\
&N_n^{(C)} \xrightarrow{w} N^{(C)} \\
&N_n^{(DCP)} \xrightarrow{w} N^{(DCP)} \quad as\ n \to \infty.
\end{aligned}
\tag{7.32}
$$

Proof. Uniform tightness of the three families of point processes is followed by Lemma 7.4. From the preliminaries in section 2, the distribution of a random measure N on $\mathcal{N}_{\mathbb{R}_+}^{\#}$ is completely determined by the finite dimensional distributions see Proposition 9.2.III in Daley and Vere-Jones (2007), i.e. the joint distribution for all finite families of bounded Borel sets $A_1, \ldots, A_k$ on $\mathbb{R}_+$ of the random variable $N(A_1), \ldots, N(A_k)$. From the tightness lemma, it is clear that all finite dimensional distribution for the point processes $N_n^{(.)}$ restricted to $[a,b]$ are uniformly tight. Consequently, there always exist a constant $T > b$ such that we can uniquely describe the finite dimensional distributions by its probability generating functional on the bounded area $[0,T]$. Combining the convergence results in Proposition 7.9, Corollary 7.4.1 and 7.4.2 in section 4, i.e. the absolute difference of the log p.g.fl.s between the point processes $N_n^{(.)}$ and $N^{(.)}$ goes to 0 as $\Delta \to 0$, equivalently $n \to \infty$

$$
\begin{aligned}
\lim_{n\to\infty} \left| \log G^{(H)}(z(.)) - \log G_{(\Delta)}^{(X_n)}(z(.)) \right| &= 0 \\
\lim_{n\to\infty} \left| \log G^{(C)}(z(.)) - \log G_{(\Delta)}^{(Y_n)}(z(.)) \right| &= 0 \\
\lim_{n\to\infty} \left| \log G^{(DCP)}(z(.)) - \log G_{(\Delta)}^{(Z_n)}(z(.)) \right| &= 0,
\end{aligned}
$$

we can now apply Proposition 7.11 and state that the families of point processes $N_n^{(H)}$, $N_n^{(C)}$ and $N_n^{(DCP)}$ converge weakly to $N^{(H)}$, $N^{(C)}$ and $N^{(DCP)}$ respectively as $n \to \infty$. □

7.6 Concluding Remarks

In this paper, we review the continuous cluster point process in a general parametric setting. Then we review the Poisson thinning INAR model and introduce the Poisson thinning INMA and the INARMA models. We prove that these integer-valued time series models, under some specific parametric setting, are actually the discrete versions of the cluster point processes $N_t^{(\cdot)}$ with continuous stochastic intensity $\lambda_t^{(\cdot)}$. We confirms Kirchner's thought in Kirchner (2016) on the relationship between the INARMA model and the dynamic contagion process. If there is a simple and effective estimation procedure for the INARMA model, for example the one Kirchner did in Kirchner (2017) for the INAR model, then the dynamic contagion process can be applied to those Hawkes-based processes. However, there are some potential issues left to be addressed. For example, can we make use of the structure standard ARMA model to perform estimation for the integer-valued version? How can we deal with random variables in the coefficients of time series models (random coefficients)? These are all proposed as topics for future research.

7.A Proof of Proposition 7.1

The Cox process is basically a cluster point process such that,

- The arrivals of cluster centres c_i follow $N^* \sim Pois(\rho)$ a homogeneous Poisson process
- Conditionally on c_i, each cluster centre will generate a cluster, the size of which follows $N_T^1 \sim Pois(\Upsilon_i f(T - c_i))$.

Vere-Jones (1970) gives the p.g.fl of a cluster process as

$$G(z(.)) = G_0(F(z(.)|t)), \tag{7.33}$$

where $G_0()$ is the p.g.fl of the process of cluster centres and $F(z(.)|t)$ is the p.g.fl for a cluster given that the cluster centre occurs at time t. Combining the second bullet point, we have

$$\begin{aligned} F^{(C)}(z(.)|c) &= \mathbb{E}[\exp \int_{\mathbb{R}_+} \log z(s) N_T^1(ds)] \\ &= \mathbb{E} \exp \left\{ \Upsilon_i \int_c^T f(s-c)(z(s)-1)ds \right\} \\ &= \mathbb{E} \exp \left\{ \Upsilon_i \int_0^{T-c} f(u)(z(c+u)-1)du \right\} \\ &= \hat{h} \left(- \int_0^{T-c} f(u)(z(c+u)-1)du \right). \end{aligned}$$

Hence the p.g.fl of the Cox process is

$$\begin{aligned} G^{(C)}(z(.)) &= \mathbb{E} \left[\exp \int_{\mathbb{R}_+} \log F^{(C)}(z(.)|c) N^*(dc) \right] \\ &= \exp \left\{ \rho \int_0^T (F^{(C)}(z(.)|c) - 1)dc \right\}. \end{aligned}$$

□

7.B Proof of Proposition 7.2

The generalized dynamic contagion process process is a cluster process,

- The arrivals of immigrants follow the Cox process with intensity $\lambda_t^{(C)}$.
- Each immigrant generates a Galton–Watson type branching process with expected branching ratio $\mu_\chi \int_0^\infty \eta(u)du < 1$. The cluster is formed by including all generations from the branching process.

Let $\mathcal{F}_t^{(C)}$ be the filtration generated by $\lambda_t^{(C)}$. Conditionally on $\mathcal{F}_t^{(C)}$, the p.g.fl of the generalized dynamic contagion process is just the p.g.fl of the Hawkes process with its immigration process being an inhomogeneous Poisson process. Then we can apply Theorem 2 in Hawkes and Oakes (1974)

$$
\begin{aligned}
G(z(.)|\mathcal{F}_t^{(C)}) &= \exp\left\{\int_0^T \left(F^{(H)}(z(.)|u) - 1\right)\lambda_u^{(c)}du\right\} \\
F^{(H)}(z(.)|u) &= z(u)\hat{g}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v) - 1)\eta(v)dv\right).
\end{aligned}
$$

The underlying intensity function is $\lambda_t = \nu + \sum_{i:\tau_i<t}\gamma(t-\tau_i)$ in Hawkes and Oakes (1974). In our case, we are working on the bounded area $[0,T]$ and $1-h(u) = 0$ when u lies outside $[0,T]$. By the definition of p.g.fl, $F(z(.)|u) = 1$ when u lies outside $[0,T]$. The ranges of integrals for $G(z(.)|\mathcal{F}_t^{(C)})$ and $F^{(H)}(z(.)|u)$, therefore, reduce to $[0,T]$ and $[0,T-u]$ respectively. Then we substitute $\gamma(t-\tau_i)$ with $\chi_i f(t-\tau_i)$ and take expectation with respect to χ_i. Finally, the unconditional p.g.fl of the generalized dynamic contagion process is $\mathbb{E}[G(z(.)|\mathcal{F}_t^{(C)})]$, which turns out to be the p.g.fl of the Cox process. Then we can apply the results from Proposition 7.1

$$
\begin{aligned}
G^{(DCP)}(z(.)) &= \mathbb{E}\left[\exp\left\{\int_0^T \left(F^{(H)}(z(.)|u) - 1\right)\lambda_u^{(C)}du\right\}\right] \\
&= \exp\left\{\rho\int_0^T\left(\hat{h}\left(-\int_0^{T-c}(F^{(H)}(z(.)|u+c) - 1)f(u)du\right) - 1\right)dc\right\}.
\end{aligned}
\tag{7.34}
$$

□

7.C Proof of Proposition 7.9

Let us define the following quantities

$$
\begin{aligned}
I_1 &= \int_0^T \Big(\hat{h}(I_2(u)) - 1\Big)\, du \\
I_2(u) &= \int_0^{T-u} (1 - F^{(H)}(z(.)|u+v))f(v)dv \\
R_1 &= \sum_{i=1}^{n} (\hat{h}(I_2((i-1)\Delta)) - 1)\Delta \\
R_2(i) &= \sum_{k=1}^{n-i+1} (1 - F^{(H)}(z(.)|(k+i-1)\Delta))f((k-1)\Delta)\Delta \\
R_3(i) &= \sum_{k=1}^{n-i+1} (1 - F^{(X_n)}(z(.)|(k+i-1)\Delta)))f((k-1)\Delta)\Delta \\
J_i &= \mu_\chi \left| \int_0^{T-i\Delta t} (F^{(H)}(z(.)|u+v) - 1)\eta(v)dv - \sum_{k=1}^{n-i} (F^{(H)}(z(.)|u+v)\eta(k\Delta)\Delta \right|.
\end{aligned}
$$

Then $D^{(DCP)}(z(.), \Delta|\Theta)$ can be decomposed as

$$
\begin{aligned}
D^{(DCP)}(z(.), \Delta|\Theta) &= \left|\log G^{(DCP)}(z(.)) - \log G^{(Z_n)}_{(\Delta)}(z(.))\right| \\
&= \rho \left| \int_0^T (\hat{h}(I_2(u)) - 1)du - \sum_{i=1}^{n} (\hat{h}(R_3(i)) - 1)\Delta \right| \\
&\leqslant \rho \left| \int_0^T (\hat{h}(I_2(u)) - 1)du - R_1 \right| + \rho \left| R_1 - \sum_{i=1}^{n} (\hat{h}(R_3(i)) - 1)\Delta \right|.
\end{aligned}
\tag{7.35}
$$

Here we add the inter-median term R_1 which is the Riemann sum of its corresponding integral. Then apply Lemma 7.1 to the first part

$$
\left| \int_0^T (\hat{h}(I_2(u)) - 1)du - R_1 \right| \sim O(\Delta^{k_1}).
$$

For the second part, we make use of the property of the convex function $\hat{h}(u)$ such that

$$\begin{aligned}\left|R_1 - \sum_{i=1}^{n}(\hat{h}(RS_3(i)) - 1)\Delta\right| &= \left|\sum_{i=1}^{n}(\hat{h}(I_2((i-1)\Delta)) - 1)\Delta - \sum_{i=1}^{n}(\hat{h}(R_3(i)) - 1)\Delta\right| \\ &\leqslant \sum_{i=1}^{n}\left|I_2((i-1)\Delta) - R_3(i)\right|\mu_\Upsilon\Delta \\ &< \sum_{i=1}^{n}\left|I_2((i-1)\Delta) - R_2(i)\right|\mu_\Upsilon\Delta + \sum_{i=1}^{n}\left|R_2(i) - R_3(i)\right|\mu_\Upsilon\Delta,\end{aligned} \tag{7.36}$$

which again separates into two parts. For the first part, apply Lemma 7.1

$$\sum_{i=1}^{n}\left|I_2((i-1)\Delta t) - R_2(i)\right|\mu_\Upsilon\Delta t \sim O(\Delta^{k_2}).$$

For the second part,

$$\begin{aligned}\left|R_2(i) - R_3(i)\right| &= \sum_{k=0}^{n-i}\left|F^{(H)}(z(.)|(k+i)\Delta) - F^{(X_n)}(z(.)|(k+i)\Delta)\right| f(k\Delta)\Delta \\ &\leqslant f_m\Delta\sum_{k=0}^{n-i}\left|F^{(H)}(z(.)|(k+i)\Delta) - F^{(X_n)}(z(.)|(k+i)\Delta)\right| \\ &\leqslant f_m\Delta\sum_{k=0}^{n}\left|F^{(H)}(z(.)|k\Delta) - F^{(X_n)}(z(.)|k\Delta)\right| \\ f_m &= \max_{k=1,\ldots,n} f(k\Delta).\end{aligned}$$

For the absolute difference $\left|F^{(H)}(z(.)|k\Delta) - F^{(X_n)}(z(.)|k\Delta)\right|$, we can solve it backwardly. When $i = n$,

$$\left|F^{(H)}(z(.)|n\Delta) - F^{(X_n)}(z(.)|n\Delta)\right| = 0.$$

When $i = n-1$,

$$
\begin{aligned}
&|F^{(H)}(z(.)|(n-1)\Delta) - F^{(X_n)}(z(.)|(n-1)\Delta)|,\\
&\leqslant z((n-1)\Delta)\left(|\hat{g}(-I_{n-1}) - \hat{g}(-R_{n-1})| + |\hat{g}(-R_{n-1}) - \hat{g}(-R^b_{n-1})|\right)\\
&\leqslant J_{n-1} + \sum_{k=1}^{1} \mu_\chi \eta(k\Delta)|F^{(H)}(z(.)|(n-1+k)\Delta) - F^{(X_n)}(z(.)|(n-1+k)\Delta)|\Delta\\
&= J_{n-1},
\end{aligned}
$$

where we use the condition $z(.) \in \mathcal{V}_0(\mathbb{R}_+)$ and $z(u) \leqslant 1, u \in (0,T)$. Then when $i = n-2$,

$$
\begin{aligned}
&|F^{(H)}(z(.)|(n-2)\Delta) - F^{(X_n)}(z(.)|(n-2)\Delta)|,\\
&\leqslant z((n-2)\Delta)\left(|\hat{g}(-I_{n-2}) - \hat{g}(-R_{n-2})| + |\hat{g}(-R_{n-2}) - \hat{g}(-R^b_{n-2})|\right)\\
&\leqslant J_{n-2} + \sum_{k=1}^{2} \mu_\chi \eta(k\Delta)|F^{(H)}(z(.)|(n-2+k)\Delta) - F^{(X_n)}(z(.)|(n-2+k)\Delta)|\Delta\\
&\leqslant J_{n-2} + J_{n-1}\mu_\chi \eta(\Delta)\Delta\\
&= J_{n-2} + O(\Delta^{k_3+1}) \sim O(\Delta^{k_3}).
\end{aligned}
$$

Note that J_i is the absolute difference between the Integral and its Riemann sum, hence we can apply Lemma 7.1

$$
\begin{aligned}
J_i &\leqslant M_i\Delta^{k_3} \leqslant M'\Delta^{k_3} \sim O(\Delta^{k_3})\\
M' &= \max_{i=0,\ldots,n} M_i.
\end{aligned}
$$

When $i = n-j$, $j = 1, 2, \ldots n$,

$$
\begin{aligned}
&|F^{(H)}(z(.)|(n-j)\Delta) - F^{(X_n)}(z(.)|(n-j)\Delta)|,\\
&\leqslant z((n-j)\Delta)\left(|\hat{g}(-I_{n-j}) - \hat{g}(-R_{n-j})| + |\hat{g}(-R_{n-j}) - \hat{g}(-R^b_{n-j})|\right)\\
&\leqslant J_{n-j} + \sum_{k=1}^{j} \mu_\chi \eta(k\Delta)|F^{(H)}(z(.)|(n-j+k)\Delta) - F^{(X_n)}(z(.)|(n-j+k)\Delta)|\Delta\\
&= J_{n-j} + jO(\Delta^{k_3+1}) \sim O(\Delta^{k_3}).
\end{aligned}
$$

Then the whole sum becomes

$$\begin{aligned}&\sum_{i=0}^{n} |F^{(H)}(z(.)|i\Delta) - F^{(X_n)}(z(.)|i\Delta)|\Delta \\ &\leqslant \Delta \sum_{i=0}^{n-1} \left(J_i + iO(\Delta^{k_3+1})\right) \sim O(\Delta^{k_3}).\end{aligned} \tag{7.37}$$

Then the second part in equation 7.36 becomes

$$\sum_{i=0}^{n} |R_2(i) - R_3(i)|\, \mu_\Upsilon \Delta \leqslant \sum_{i=1}^{n} \left(f_m \Delta \sum_{k=0}^{n} \left|F^{(H)}(z(.)|k\Delta) - F^{(X_n)}(z(.)|k\Delta)\right| \right) \mu_\Upsilon \Delta \sim O(\Delta^{k_3}).$$

Finally, let $k = \min\{k_1, k_2, k_3\}$. □

7.D Proof of Corollary 7.4.1

The results follows from Proposition 7.9. The p.g.fl of the Cox process $G^{(C)}$ can be derived from the p.g.fl of the generalized dynamic contagion process $G^{(DCP)}$ by letting $\eta(u) = 0$ such that $F^{(H)}(z(.))$ becomes

$$F^{(H)}(z(.)|u) = z(u)\hat{g}(0) = z(u).$$

Similarly, $G^{(Y_n)}_{(\Delta)}$ can be derived from $G^{(Z_n)}_{(\Delta)}$ by letting $F^{(X_n)}(z(.)|i\Delta) = z(i\Delta)$. Then $D^{(C)}(z(.), \Delta|\Theta)$ will have the same form as the equations 7.35 and 7.36 such that

$$\begin{aligned}D^{(C)}(z(.), \Delta|\Theta) &= \left|\log G^{(C)}(z(.)) - \log G^{(Y_n)}_{(\Delta))}(z(.))\right| \\ &= \rho \left| \int_0^T (\hat{h}(I_2(u)) - 1)du - \sum_{i=1}^{n} (\hat{h}(R_3(i)) - 1)\Delta \right| \\ &\leqslant \rho \left| \int_0^T (\hat{h}(I_2(u)) - 1)du - R_1 \right| + \rho \left| R_1 - \sum_{i=1}^{n} (\hat{h}(R_3(i)) - 1)\Delta \right| \\ &\leqslant \rho \left| \int_0^T (\hat{h}(I_2(u)) - 1)du - R_1 \right| + \rho \sum_{i=1}^{n} |I_2((i-1)\Delta t) - R_2(i)|\, \mu_\Upsilon \Delta \\ &\quad + \rho \sum_{i=1}^{n} |R_2(i) - R_3(i)|\, \mu_\Upsilon \Delta \\ &\sim O(\Delta^{k_1}) + O(\Delta^{k_2}),\end{aligned}$$

where $R_2(i) - R_3(i) = 0$ since $F^{(H)}(z(.)|i\Delta) = z(i\Delta) = F^{(X_n)}(z(.)|i\Delta)$. Finally, we can take $k = \min\{k_1, k_2\}$. □

7.E Proof of Corollary 7.4.2

Similarly, this result follows from Proposition 7.9. From the p.g.fl.s point of view, $G^{(H)}$ can be recovered by replacing ρ and $\hat{h}\left(-\int_0^{T-u}(F^{(H)}(z(.)|u+v)-1)f(v)dv\right)$ by ν and $F^{(H)}(z(.)|u)$ in $G^{(DCP)}$ respectively. $G_{(\Delta)}^{(X_n)}$ can be derived from $G_{(\Delta)}^{(Z_n)}$ in a similar way. Then $D^{(H)}$ becomes

$$\begin{aligned} D^{(H)}(z(.),\Delta|\Theta) &= \nu\left|\int_0^T (F^{(H)}(z(.)|u))-1)du - \sum_{i=1}^{n}(F^{(X_n)}(z(.)|i\Delta)-1)\Delta\right| \\ &\leqslant \nu\left|\int_0^T (F^{(H)}(z(.)|u)-1)du - \sum_{i=1}^{n}(F^{(H)}(z(.)|i\Delta)-1)\Delta\right| \\ &+ \nu\left|\sum_{i=1}^{n}(F^{(H)}(z(.)|i\Delta)-1)\Delta - \sum_{i=1}^{n}(F^{(X_n)}(z(.)|i\Delta)-1)\Delta\right|. \end{aligned}$$

Adopting the similar technique as in Proposition 7.9, we add the term $\sum_{i=1}^{n}(F^{(H)}(z(.)|i\Delta)-1)\Delta$ which is the right Riemann sum of the integral. Then we can apply Lemma 7.1

$$\left|\int_0^T (F^{(H)}(z(.)|u)-1)du - \sum_{i=1}^{n}(F^{(H)}(z(.)|i\Delta)-1)\Delta\right| \sim O(\Delta^{k_1}), \quad k_1 > 0.$$

The second part is

$$\begin{aligned} &\left|\sum_{i=1}^{n}(F^{(H)}(z(.)|i\Delta)-1)\Delta - \sum_{i=1}^{n}(F^{(X_n)}(z(.)|i\Delta)-1)\Delta\right| \\ &\leqslant \sum_{i=1}^{n}|F^{(H)}(z(.)|i\Delta) - F^{(X_n)}(z(.)|i\Delta)|\Delta \sim O(\Delta^{k_2}). \end{aligned}$$

This result follows from the inequality 7.37. Finally, we can take $k = \min\{k_1, k_2\}$. □

CHAPTER 8

Paper E. INAR Approximation of Bivariate Linear Birth and Death Process

Abstract

In this paper, we propose a new type of univariate and bivariate Integer-valued autoregressive model of order one (INAR(1)) to approximate univariate and bivariate linear birth and death process with constant rates. Under a specific parametric setting, the dynamic of transition probabilities and probability generating function of INAR(1) will converge to that of birth and death process as the length of subintervals goes to 0. Due to the simplicity of Markov structure, maximum likelihood estimation is feasible for INAR(1) model, which is not the case for bivariate and multivariate birth and death process. This means that the statistical inference of bivariate birth and death process can be achieved via the maximum likelihood estimation of a bivariate INAR(1) model.

8.1 Introduction

The simple linear birth and death process, which was first introduced by Feller (1939), is a widely used Markov model with applications in population growth, epidemiology, genetics and so on. The basic idea of this process is that the probabilities

of any individual giving birth to a new individual, or any individual dying, are constant at any moment in time and all individuals are independent of each other. Many statistical properties, including moments, distribution function, extinction probability, or some other cumulative distribution of interests, are explicitly derived in the literature; see for example, Kendall (1949). The statistical inference for simple birth and death processes is then developed by Keiding (1975), where maximum likelihood estimators and other asymptotic results are discussed. Since the distribution function of simple birth and death processes is explicit, the construction of the likelihood function is straightforward. However, it is pointed out in the literature that the transition probability is actually cumbersome and numerically unstable when the size of population is large over time. At the same time, a variety of alternative estimation methods have been proposed. For example, quasi- and pseudo – likelihood estimators Chen and Hyrien (2011), Crawford et al. (2014) addressed it as a missing data problem and apply an EM algorithm to maximize it. Tavaré (2018) found those transition probabilities by numerical inversion of the probability generating function and then applied Bayesian methods to perform estimation. Davison et al. (2021) adopted a saddle point approximation method to further improve the accuracy of transition probabilities.

The bivariate and multivariate birth and death process are developed in Griffiths (1972, 1973). Griffiths (1972) described the transmission of malaria (so called host-vector situation) as a bivariate birth and death process where there is no direct infection between the same type of population. Then the author extended the model to multivariate case Griffiths (1973) which can be regarded as an approximation of general epidemic with several types of infective. However, due to the intractability of the joint probability generating function, maximum likelihood estimation for parameters is not implementable. One possible way forward is to use integer-valued time series to approximate the continuous birth and death process and maximum likelihood estimation would then be feasible.

In recent years, there has been a growing interest in modelling integer-valued time series due to the presence of count data from different scientific fields such as social science, healthcare, insurance, economic and the financial industry. In particular, regarding to the univariate case, Al-Osh and Alzaid (1987) and McKenzie (1985) were the first to consider an INAR(1) model based on the so-called binomial thinning operator. The idea here is to manipulate the operation between coefficients and variables as well as the innovation terms in a way that the values are always integer. One can apply different discrete random variables to describe this operation. For more details, the interested reader can refer to Weiß (2018) Davis et al. (2016a), Scotto et al. (2015), Weiß (2008b) among many more.

In this paper, we propose an integer-valued autoregressive model of order one (INAR(1)) to approximate continuous birth and death process. In this way, the continuous process is approximated by a discrete Markov chain so that transition probabilities as well as likelihood function can be written down explicitly. As the birth and death process in our setting does not consider any immigrant, the innovation term is dropped in the proposed INAR(1) model. Similar to Nelson (1990), Kirchner (2016), where they find out the relationship between discrete models and their continuous counterparts, we also first need to make sure the our proposed discrete INAR(1) model would converge to birth and death process in weak convergence sense. Then we will explore how our proposed model would help facilitate the statistical inference. According to the probability generating function of the simple birth and death process, the death part can be described by binomial random variable while the birth part corresponds to a negative binomial. Then one can construct a bivariate INAR model based on these random variables to describe the bivariate birth and death process and even the multivariate one. As the transition probabilities and likelihood function of bivariate birth and death process cannot be written down explicitly, the main contribution is that the proposed bivariate INAR(1) model would provide a feasible way to estimate the parameters of bivariate birth and death process (Maximum likelihood estimation).

The paper is organized as follows: Section 2 reviews some main results of univariate and bivariate birth and death processes with constant rates. Section 3 introduces Integer-valued autoregressive models as well as some distributional properties. Section 4 constructs the discrete semimartingale using the proposed INAR models and proves the weak convergence between constructed semimartingale and birth and death processes. A simulation study is carried out in section 5 to illustrate the estimation method via proposed INAR models an their corresponding properties of estimators. Some concluding remarks are in section 6.

8.2 Univariate and bivariate birth and death processes

In this section, we will review the essential elements of simple birth-and-death processes, including moments and other distributional properties. These are well known and extensively discussed in the literature. Then, we will discuss the bivariate case where analytic expressions of the distribution function are not available.

8.2.1 Simple univariate birth-and-death process

Suppose that we have a population whose total number is evolved as a simple birth and death process Z_t, with constant birth rate $\lambda \geqslant 0$, death rate $\mu \geqslant 0$ and initial population $Z_0 \in \mathbb{N}$. In other words, the probability that any individual gives birth in time Δ is $\lambda\Delta$, and the probability that any individual dies in time Δ is $\mu\Delta$. Individuals are independent of each other. Let $P_n(t) = \Pr(Z_t = n)$ be the probability that the total population is n at time t. Then the transition probability of the simple birth and death process is characterized by the following ordinary differential equation (ODE)

$$\begin{cases} \frac{dP_n(t)}{dt} & = \lambda(n-1)P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - (\lambda+\mu)nP_n(t), \quad n \geqslant 1 \\ P_{Z_0}(0) & = 1 \end{cases} \tag{8.1}$$

Applying a liner transform $\sum_n \theta^n$ on both sides and $\varphi(t,\theta) = \sum_n \theta^n P_n(t)$, we can get a partial differential equation whose solution φ is the probability generating function of $Z_t^{(a)}$.

$$\begin{aligned} \frac{\partial\varphi}{\partial t} &= \lambda\theta^2\frac{\partial\varphi}{\partial\theta} + \mu\frac{\partial\varphi}{\partial\theta} - (\lambda+\mu)\theta\frac{\partial\varphi}{\partial\theta} \\ &= (\lambda\theta-\mu)(\theta-1)\frac{\partial\varphi}{\partial\theta} \\ \varphi(0,\theta) &= \theta^a \end{aligned} \tag{8.2}$$

This linear PDE can be solved explicitly

$$\begin{aligned} \varphi(t,\theta) &= \left(1-\alpha(t)+\alpha(t)\frac{\beta(t)\theta}{1-(1-\beta(t))\theta}\right)^{Z_0} \\ \alpha(t) &= \frac{(\lambda-\mu)e^{(\lambda-\mu)t}}{\lambda e^{(\lambda-\mu)t}-\mu}, \quad \beta(t) = \frac{\lambda-\mu}{\lambda e^{(\lambda-\mu)t}-\mu} \end{aligned} \tag{8.3}$$

This probability generating function clearly gives the construction of Z_t given Z_0, i.e. the sum of i.i.d zero-modified geometric random variables

$$Z_t \sim \sum_{i=1}^{Z_0} B_i(\alpha(t))G_i(\beta(t)), \tag{8.4}$$

where B_i are i.i.d Bernoulli random variables and G_i are i.i.d Geometric random variables with mean $\alpha(t)$ and $\frac{1}{\beta(t)}$, respectively. Furthermore, from the definition of

transition probability, the linear birth and death process is a pure-jump semimartingale with following characteristic triplet:

$$Ch(Z_t) = \begin{cases} B_t = 0 \\ C_t = 0 \\ \nu(Z_t; dt, dx) = dtK(Z_t, dx) = dt(\lambda Z_{t^-}\delta_1(dx) + \mu Z_{t^-}\delta_{-1}(dx)) \end{cases} \tag{8.5}$$

$$\int_R \left(x^2 \wedge 1\right) K(Z_t, dx) = (\lambda + \mu) Z_{t^-} < \infty, \quad \text{given that } Z_{t^-} \text{ is finite}$$

With the help of piece-wise deterministic Markov process theory in Davis (1984), the infinitesimal generator of the simple birth and death process Z_t acting on a function $f(t, Z)$ within its domain $\Omega(\mathcal{A})$ is given by

$$\mathcal{A}f(t, Z) = \frac{\partial f}{\partial t} + \lambda Z(f(t, Z+1) - f(t, Z)) + \mu Z(f(t, Z-1) - f(t, Z)), \tag{8.6}$$

where $\Omega(\mathcal{A})$ is the domain for the generator $\mathcal{A}$ such that $f(t, Z)$ is differentiable with respect to t for all t, Z, and

$$\begin{aligned} |f(t, Z+1) - f(t, Z)| < \infty \\ |f(t, Z-1) - f(t, Z)| < \infty. \end{aligned} \tag{8.7}$$

The first and second moments can be derived by applying infinitesimal generator to the functions $f(t, z) = Z, Z^2$ such that

$$\begin{aligned} \mathcal{A}Z &= \lambda Z(Z + 1 - Z) + \mu Z(Z - 1 - Z) \\ \mathcal{A}Z^2 &= \lambda Z((Z+1)^2 - Z^2) + \mu Z((Z-1)^2 - Z^2), \end{aligned} \tag{8.8}$$

which leads to two ODEs,

$$\begin{aligned} \frac{d\mathbb{E}[Z_t]}{dt} &= (\lambda - \mu)\mathbb{E}[Z_t] \\ \frac{d\mathbb{E}[Z_t^2]}{dt} &= 2(\lambda - \mu)\mathbb{E}[Z_t^2] + (\lambda + \mu)\mathbb{E}[Z_t] \end{aligned} \tag{8.9}$$

Then, we can solve them explicitly

$$\begin{aligned}
\mathbb{E}[Z_t] &= Z_0 e^{(\lambda-\mu)t} \\
\mathbb{E}[Z_t^2] &= Z_0^2 e^{2(\lambda-\mu)t} + \frac{Z_0(\lambda+\mu)}{(\lambda-\mu)} e^{(\lambda-\mu)t}\left(e^{(\lambda-\mu)t}-1\right) \\
Var(Z_t) &= \frac{Z_0(\lambda+\mu)}{(\lambda-\mu)} e^{(\lambda-\mu)t}\left(e^{(\lambda-\mu)t}-1\right)
\end{aligned} \tag{8.10}$$

According to the analytic expression of the first moment, it is clear that the population is bound to become extinct if $\lambda < \mu$.

8.2.2 Bivariate birth-and-death process

Suppose there are two populations $\mathbf{M} = (M_1, M_2)^T$ with initial population $\mathbf{M}_0 \in \mathbb{N}_+^2$. The rate with which the population M_1 increases by one is $\lambda_{21}M_2 + \lambda_{11}M_1$ while the same for the population M_2 would be $\lambda_{12}M_1 + \lambda_{22}M_2$. The subscript $\lambda_{i,j}$ means that the rate is from population i contributed to population j. The death rate for two populations would be μ_1, μ_2 respectively. The two population is not independent as long as the cross birth rates $\lambda_{i,j} \neq 0,\ i \neq j$. Then denote $P_{mn}(t) = \Pr(M_{1,t} = m, M_{2,t} = n)$. This satisfies the following ODE

$$\begin{cases}
\frac{dP_{m,n}}{dt} = & (\lambda_{11}(m-1) + \lambda_{21}n)\, P_{m-1,n} + \mu_1(m+1)P_{m+1,n} \\
& + (\lambda_{12}m + \lambda_{22}(n-1))\, P_{m,n-1} + \mu_2(n+1)P_{m,n+1} \\
& - ((\lambda_{11}+\lambda_{12}+\mu_1)m + (\lambda_{21}+\lambda_{22}+\mu_2)n)\, P_{m,n} \\
P_{\mathbf{M}_0}(0) = & 1, \quad M_{1,0}, M_{2,0} \in \mathbb{N}_+
\end{cases} \tag{8.11}$$

Griffiths (1972) introduced this bivariate birth death process ($\lambda_{11} = \lambda_{22} = 0$) to describe the host-vector epidemic situation where the birth probability of two population depends on the size the other population only, e.g. transmission of malaria. To get the joint probability generating function of $\Psi(t, \theta, \phi) = \sum_m \sum_n \theta^m \phi^n P_{mn}(t)$, we can apply a linear transform $\sum_m \sum_n \theta^m \phi^n$ on both sides of the ODE. The resulting PDE is

$$
\begin{aligned}
\frac{\partial \Psi}{\partial t} &= \lambda_{11}\theta^2\frac{\partial \Psi}{\partial \theta} + \lambda_{21}\theta\phi\frac{\partial \Psi}{\partial \phi} + \mu_1\frac{\partial \Psi}{\partial \theta} + \lambda_{12}\theta\phi\frac{\partial \Psi}{\partial \theta} + \lambda_{22}\phi^2\frac{\partial \Psi}{\partial \phi} + \mu_2\frac{\partial \Psi}{\partial \phi} \\
&\quad - \theta(\lambda_{11} + \lambda_{12} + \mu_1)\frac{\partial \Psi}{\partial \theta} - \phi(\lambda_{21} + \lambda_{22} + \mu_2)\frac{\partial \Psi}{\partial \phi} \\
&= (\lambda_{11}\theta^2 + \lambda_{12}\theta\phi + \mu_1 - \theta(\lambda_{11} + \lambda_{12} + \mu_1))\frac{\partial \Psi}{\partial \theta} \\
&\quad + (\lambda_{22}\phi^2 + \lambda_{21}\theta\phi + \mu_2 - \phi(\lambda_{21} + \lambda_{22} + \mu_2))\frac{\partial \Psi}{\partial \phi} \\
\Psi(0, \theta, \phi) &= \theta^{M_{1,0}}\phi^{M_{2,0}}
\end{aligned}
\tag{8.12}
$$

This is a semi-linear PDE. The subsidiary equations are defined as

$$
\begin{aligned}
\frac{d\Psi}{0} = \frac{dt}{1} &= \frac{-d\theta}{\lambda_{11}\theta^2 + \lambda_{12}\theta\phi + \mu_1 - \theta(\lambda_{11} + \lambda_{12} + \mu_1)} \\
&= \frac{-d\phi}{\lambda_{22}\phi^2 + \lambda_{21}\theta\phi + \mu_2 - \phi(\lambda_{21} + \lambda_{22} + \mu_2)}
\end{aligned}
\tag{8.13}
$$

The first fraction does not mean divide $d\Psi$ by 0 and combining with the second fraction $\frac{dt}{1}$ infers that Ψ = constant, according to chapter 8 of Bailey (1991). Matching the third and fourth differentials above, we have

$$
\frac{d\theta}{d\phi} = \frac{\lambda_{11}\theta^2 + \lambda_{12}\theta\phi + \mu_1 - \theta(\lambda_{11} + \lambda_{12} + \mu_1)}{\lambda_{22}\phi^2 + \lambda_{21}\theta\phi + \mu_2 - \phi(\lambda_{21} + \lambda_{22} + \mu_2)} \tag{8.14}
$$

It seems that there is no way to solve this non-linear ODE and therefore no explicit solution is available for this PDE. However, it can be shown that this PDE gives a unique solution by Existence-Uniqueness Theorem for Quasilinear First-Order Equations. With regard to its characteristic, similar to the univariate case, this

process is a pure-jump semimartingale with following characteristic triplets:

$$Ch(\mathbf{M}_t) = \begin{cases} B_t = 0 \\ C_t = 0 \\ \nu(\mathbf{M}_t; dt, dx) = dtK(\mathbf{M}_t, dx) = \\ dt(\tilde{\boldsymbol{\lambda}}_1\delta_{(1,0)}(dx) + \tilde{\boldsymbol{\lambda}}_2\delta_{(0,1)}(dx) + \tilde{\boldsymbol{\mu}}_1\delta_{(-1,0)}(dx) + \tilde{\boldsymbol{\mu}}_2\delta_{(0,-1)}(dx))\mathbf{M}_{t^-} \end{cases}$$

$$\int_R \left(x^2 \wedge 1\right) K(M_t, dx) = (\tilde{\boldsymbol{\lambda}}_1 + \tilde{\boldsymbol{\lambda}}_2 + \tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\mu}}_2)\mathbf{M}_{t^-} < \infty, \text{ given that } \mathbf{M}_{t^-} \text{ is finite,}$$

where

$$\tilde{\boldsymbol{\lambda}}_1 = (\lambda_{11}, \lambda_{21}), \quad \tilde{\boldsymbol{\lambda}}_2 = (\lambda_{21}, \lambda_{22}), \quad \tilde{\boldsymbol{\mu}}_1 = (\mu_1, 0), \quad \tilde{\boldsymbol{\mu}}_2 = (0, \mu_2) \tag{8.15}$$

The moments of this bivariate process can be derived by applying again infinitesimal generator.

Proposition 8.1. *The first and second moments of the bivariate birth and death process* $\mathbf{M}_t = (M_{1,t}, M_{2,t})$ *defined in (8.11) are given by*

$$\begin{aligned} \mathbb{E}[M_{1,t}] &= M_{1,0}\left(\frac{\lambda_{12}c}{2\lambda_{12}c + \kappa_1 - \kappa_2}e^{(\lambda_{12}c-\kappa_2)t} + \frac{\lambda_{12}c + \kappa_1 - \kappa_2}{2\lambda_{12}c + \kappa_1 - \kappa_2}e^{-(\lambda_{12}c+\kappa_1)t}\right) \\ &\quad + M_{2,0}\frac{\lambda_{21}}{2\lambda_{12}c + \kappa_1 - \kappa_2}\left(e^{(\lambda_{12}c-\kappa_2)t} - e^{-(\lambda_{12}c+\kappa_1)t}\right) \\ \mathbb{E}[M_2, t] &= M_{1,0}\frac{\lambda_{12}}{2\lambda_{12}c + \kappa_1 - \kappa_2}\left(e^{(\lambda_{12}c-\kappa_2)t} - e^{-(\lambda_{12}c+\kappa_1)t}\right) \\ &\quad + M_{2,0}\left(\frac{\lambda_{12}c + \kappa_1 - \kappa_2}{2\lambda_{12}c + \kappa_1 - \kappa_2}e^{(\lambda_{12}c-\kappa_2)t} + \frac{\lambda_{12}c}{2\lambda_{12}c + \kappa_1 - \kappa_2}e^{-(\lambda_{12}c+\kappa_1)t}\right), \end{aligned} \tag{8.16}$$

where

$$\kappa_1 = \mu_1 - \lambda_{11}, \quad \kappa_2 = \mu_2 - \lambda_{22}, \quad c = \frac{\kappa_2 - \kappa_1 + \sqrt{(\kappa_1 - \kappa_2)^2 + 4\lambda_{21}\lambda_{12}}}{2\lambda_{12}}.$$

The second moments $\mathbb{E}[M_{1,t}^2], \mathbb{E}[M_{2,t}^2]$ *and* $\mathbb{E}[M_{1,t}M_{2,t}]$ *are determined by the follow-*

ing system of ODE,

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[M_{1,t}^2] &= -2\kappa_1\mathbb{E}[M_{1,t}^2] + 2\lambda_{21}\mathbb{E}[M_{1,t}M_{2,t}] + \lambda_{21}\mathbb{E}[M_{2,t}] + \mu_1\mathbb{E}[M_{1,t}] \\
\frac{d}{dt}\mathbb{E}[M_{2,t}^2] &= -2\kappa_2\mathbb{E}[M_{2,t}^2] + 2\lambda_{12}\mathbb{E}[M_{1,t}M_{2,t}] + \lambda_{12}\mathbb{E}[M_{1,t}] + \mu_2\mathbb{E}[M_{2,t}] \\
\frac{d}{dt}\mathbb{E}[M_{1,t}M_{2,t}] &= -(\kappa_1+\kappa_2)\mathbb{E}[M_{1,t}M_{2,t}] + \lambda_{21}\mathbb{E}[M_{2,t}^2] + \lambda_{12}\mathbb{E}[M_{1,t}^2]
\end{aligned} \tag{8.17}$$

Proof. See appendix A.8.A □

Note that to ensure the bivariate process becomes extinct with probability one, we need the (necessary and sufficient condition) $(\mu_1 - \lambda_{11})(\mu_2 - \lambda_{22}) > \lambda_{12}\lambda_{21}$ according to Griffiths (1973). Many interesting properties of the process have been investigated by Griffiths (1972, 1973). In general, this bivariate birth and death process is not straightforward to apply in practice because there are no explicit solutions to the above PDE, and the second moments have to be evaluated by numerical methods. The discrete integer-value model proposed in the next section would be a possible solution.

8.3 Univariate and Bivariate INAR models

In this section, we will introduce integer-valued autoregressive models which will serve as discrete approximations for continuous counterparts discussed in the last section. The derivation of this approximation will demonstrate how to parameterize the bivariate INAR case.

8.3.1 Univariate INAR model

The classical integer-value autoregressive (INAR) model is introduced by defining a so-called binomial thinning operator $\circ$ such that $\alpha \circ X$ is the sum of X i.i.d Bernoulli random variable with success probability α. i.e.

$$\alpha \circ X = \sum_{i=1}^{X} b_i, \quad b_i \overset{i.i.d}{\sim} \text{Bernoulli}(\alpha) \tag{8.18}$$

A well-known Poisson INAR(1) model X_t is given by

$$X_t = \alpha \circ X_{t-1} + R_t, \tag{8.19}$$

where $\{R_i\}_{i=1,\dots,t}$ are i.i.d Poisson variables with parameter ρ. The key idea of the integer-value model would be the operator $\circ$. One can choose different discrete random variables to construct different integer-valued models. Indicated by the transition probability of continuous birth and death process, i.e. the sum of i.i.d zero-modified geometric random variables shown in equation (8.4), INAR model can be a good approximation by combining $\circ$ and geometric operator as defined below.

Definition 8.1. *A birth and death INAR(1) model with survival probability* $\alpha \in [0,1]$ *and birth probability* $p \in [0,1]$ *is defined as*

$$X_t = p *_1 \alpha \circ X_{t-1}, \tag{8.20}$$

where

- $\circ$ *is the binomial operator*
- $*_1$ *is a geometric (reproduction) operator such that* $p *_1 X = \sum_{i=1}^{X} g_i^{(1)}$ *with* $g_i^{(1)}$ *being i.i.d geometric random variable with success probability* p *whose probability mass function is given by*

$$P(g_i^{(1)} = k) = p(1-p)^{k-1}, \quad k = 1, 2, \dots,$$

- $p *_1 \alpha \circ X = \sum_{i=1}^{\alpha \circ X} g_i^{(1)}$

Remark The innovation is dropped as there is no independent immigrant process in the birth and death process investigated.

Proposition 8.2. *The birth and death INAR(1) model has the following statistical properties*

1. *The probability generating function of* X_t *can be iterated backwardly such that*

$$\begin{aligned} \varphi^{(I)}(t,\theta) = \mathbb{E}[\theta^{X_t}] &= \mathbb{E}\left[\left(1 - \alpha + \frac{\alpha p \theta}{1-(1-p)\theta}\right)^{X_{t-1}}\right] \\ &= \mathbb{E}\left[\left(1 - \alpha_i + \frac{\alpha_i p_i \theta}{1-(1-p_i)\theta}\right)^{X_{t-i}}\right], \quad i = 1, \dots, t \end{aligned} \tag{8.21}$$

where

$$
\begin{aligned}
p_i &= \frac{p^i}{d_{i-1}} \quad \alpha_i = \frac{\alpha^i}{d_{i-1}} \\
d_i &= p^i \left(1 + (1-p) \frac{\frac{\alpha}{p} - \left(\frac{\alpha}{p}\right)^{i+1}}{1 - \frac{\alpha}{p}} \right)
\end{aligned} \tag{8.22}
$$

In order words, the birth and death operator $p *_1 \alpha\circ$ as a whole is iterable.

$$
X_t = p_1 *_1 \alpha_1 \circ X_{t-1} = p_2 *_1 \alpha_2 \circ X_{t-2} = \cdots = p_t *_1 \alpha_t \circ X_0 \tag{8.23}
$$

2. *Then the mean, variance and covariance are given by*

$$
\begin{aligned}
\mathbb{E}[X_t] &= \frac{\alpha_i}{p_i} \mathbb{E}[X_{t-i}] \\
Var(X_t) &= \left(\frac{\alpha_i(1-p_i)}{p_i^2} + \frac{\alpha_i(1-\alpha_i)}{p_i^2} \right) \mathbb{E}[X_{t-i}] + \frac{\alpha_i^2}{p_i^2} Var(X_{t-i}) \\
Cov(X_t, X_{t-i}) &= \frac{\alpha_i}{p_i} Var(X_{t-i})
\end{aligned} \tag{8.24}
$$

Proof. See appendix A.8.B. □

Note that if $\alpha/p < 1$, the process X_t will become extinct eventually. It is obvious that the continuous birth and death process can be approximated by this discrete INAR(1) model by directly matching the probability generating function $\varphi^{(I)}$ to the one φ in equation (8.3) as the $p *_1 \alpha \circ X$ is the sum of X i.i.d zero-modified geometric random variables.

8.3.2 Bivariate INAR model

Discrete approximation for univariate birth and death process is somehow simple because the PDE(8.2) has an explicit solution and hence the distribution is already known. In the case where the dynamic of two populations are characterized by (8.11), no explicit solution for its PDE (8.12). However, from the birth and death INAR(1) model, it is clear that birth and death probability are closely related to binomial and negative binomial random variables. Based on the dynamic (8.11) and linear form of the first moment (8.16), a bivariate INAR(1) model is proposed as follows.

Definition 8.2. *A bivariate birth and death INAR(1) model* $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})^T$ *with survival probability* $\alpha_1, \alpha_2 \in [0,1]$ *and birth probability* $\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22} \in [0,1]$ *is defined as*

$$\begin{aligned} Y_{1,t} &= \beta_{11} *_1 \alpha_1 \circ Y_{1,t-1} + \beta_{21} *_2 Y_{2,t-1} \\ Y_{2,t} &= \beta_{12} *_2 Y_{1,t-1} + \beta_{22} *_1 \alpha_2 \circ Y_{2,t-1}, \end{aligned} \tag{8.25}$$

where

- $\circ$ *is the binomial operator*
- $*_2$ *is another geometric (reproduction) operator different from* $*_1$ *such that* $\beta *_2 X = \sum_{i=1}^{X} g_i^{(2)}$ *with* $g_i^{(2)}$ *being i.i.d geometric random variable whose success probability is* β *. The probability mass function is given by*

$$P(g_i^{(2)} = k) = \beta(1-\beta)^k, \quad k = 0, 1, 2, \ldots,$$

- *Conditional on* $\mathbf{Y}_{t-1}$*, the random variables* $\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}$*,* $\beta_{21} *_2 Y_{2,t-1}$*,* $\beta_{12} *_2 Y_{1,t-1}$ *and* $\beta_{22} *_1 \alpha_2 \circ Y_{1,t-1}$ *are all independent of each other.*

Now it seems that the structure of bivariate INAR(1) matches the the dynamics of (8.11), i.e. the birth probability depends on the size of both populations while death probability depends on the size of its own population. We adopt another geometric random variable $g^{(2)}$ which is slightly different from $g^{(1)}$ because for example, if we use $g^{(1)}$, $Y_{1,t} \geqslant Y_{2,t-1} \forall t$ which is not reasonable when $Y_{1,t-1} < Y_{2,t-1}$ for a population.

Proposition 8.3. *The first and second moments of the bivariate INAR(1) defined*

above are characterized by the following recursive formulas

$$
\begin{aligned}
\mathbb{E}[Y_{1,t}] &= \frac{\alpha_1}{\beta_{11}}\mathbb{E}[Y_{1,t-1}] + \frac{1-\beta_{21}}{\beta_{21}}\mathbb{E}[Y_{2,t-1}] \\
\mathbb{E}[Y_{2,t}] &= \frac{1-\beta_{12}}{\beta_{12}}\mathbb{E}[Y_{1,t-1}] + \frac{\alpha_2}{\beta_{22}}\mathbb{E}[Y_{2,t-1}] \\
Var(Y_{1,t}) &= \frac{\alpha_1^2}{\beta_{11}^2}Var(Y_{1,t-1}) + \frac{\alpha_1(2-\beta_{11}-\alpha_1)}{\beta_{11}^2}\mathbb{E}[Y_{1,t-1}] + \left(\frac{1-\beta_{21}}{\beta_{21}}\right)^2 Var(Y_{2,t-1}) \\
&\quad + \frac{1-\beta_{21}}{\beta_{21}^2}\mathbb{E}[Y_{2,t-1}] + 2\frac{\alpha_1(1-\beta_{21})}{\beta_{11}\beta_{21}}Cov(Y_{1,t-1}, Y_{2,t-1}) \\
Var(Y_{2,t}) &= \left(\frac{1-\beta_{12}}{\beta_{12}}\right)^2 Var(Y_{1,t-1}) + \frac{1-\beta_{12}}{\beta_{12}^2}\mathbb{E}[Y_{1,t-1}] + \frac{\alpha_2^2}{\beta_{22}^2}Var(Y_{2,t-1}) \\
&\quad + \frac{\alpha_2(2-\beta_{22}-\alpha_2)}{\beta_{22}^2}\mathbb{E}[Y_{2,t-1}] + 2\frac{\alpha_2(1-\beta_{12})}{\beta_{12}\beta_{22}}Cov(Y_{1,t-1}, Y_{2,t-1}) \\
Cov(Y_{1,t}, Y_{2,t}) &= \left(\frac{\alpha_1\alpha_2}{\beta_{11}\beta_{22}} + \frac{(1-\beta_{21})(1-\beta_{12})}{\beta_{12}\beta_{21}}\right) Cov(Y_{1,t-1}, Y_{2,t-1}) \\
&\quad + \frac{\alpha_1(1-\beta_{12})}{\beta_{11}\beta_{12}}Var(Y_{1,t-1}) + \frac{\alpha_2(1-\beta_{21})}{\beta_{21}\beta_{22}}Var(Y_{2,t-1})
\end{aligned}
\tag{8.26}
$$

Proof. Similar to proposition 8.2, the moments can be derived by conditional expectation. The first and second moment for random variable $g_i^{(2)}$ with parameter β are $\frac{1-\beta}{\beta}$ and $\frac{1-\beta}{\beta^2}$. Then the first moment for X_t are

$$
\begin{aligned}
\mathbb{E}[Y_{1,t}|\mathbf{Y}_{t-1}] &= \mathbb{E}[\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}|Y_{1,t-1}] + \mathbb{E}[\beta_{21} *_2 Y_{2,t-1}|Y_{2,t-1}] \\
&= \frac{\alpha_1}{\beta_{11}}Y_{1,t-1} + \frac{1-\beta_{21}}{\beta_{21}}Y_{2,t-1}
\end{aligned}
$$

The second moments are given by

$$
\begin{aligned}
Var(Y_{1,T}|\mathbf{Y}_t) &= Var(\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}|Y_{1,t-1}) + Var(\beta_{21} *_2 Y_{2,t-1}|Y_{2,t-1}) \\
&= \frac{\alpha_1(2-\beta_{11}-\alpha_1)}{\beta_{11}^2}Y_{1,t-1} + \frac{1-\beta_{21}}{\beta_{21}^2}Y_{2,t-1} \\
Var(Y_{1,t}) &= Var(\mathbb{E}[Y_{1,t-1}|\mathbf{Y}_{t-1}]) + \mathbb{E}[Var(Y_{1,t}|\mathbf{Y}_{t-1})] \\
&\quad + 2Cov(\mathbb{E}[\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}|Y_{1,t-1}], \mathbb{E}[\beta_{21} *_2 Y_{2,t-1}|Y_{2,t-1}]) \\
&\quad + 2\mathbb{E}[Cov(\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}, \beta_{21} *_2 Y_{2,t-1}|\mathbf{Y}_{t-1})] \\
&= Var(\mathbb{E}[Y_{1,t}|\mathbf{Y}_{t-1}]) + \mathbb{E}[Var(Y_{1,t}|\mathbf{Y}_{t-1})] \\
&\quad + \frac{\alpha_1(1-\beta_{21})}{\beta_{11}\beta_{21}}Cov(Y_{1,t-1}, Y_{2,t-1})
\end{aligned}
$$

$$
\begin{aligned}
Cov(Y_{1,t}, Y_{2,t}) &= Cov(\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}, \beta_{12} *_2 Y_{1,t-1}) + Cov(\beta_{11} *_1 \alpha_1 \circ Y_{1,t-1}, \beta_{22} *_1 \alpha_2 \circ Y_{2,t-1}) \\
&+ Cov(\beta_{21} *_2 Y_{2,t-1}, \beta_{12} *_2 Y_{1,t-1}) + Cov(\beta_{21} *_2 Y_{2,t-1}, \beta_{22} *_1 \alpha_2 \circ Y_{2,t-1}) \\
&= \frac{\alpha_1(1-\beta_{12})}{\beta_{11}\beta_{12}} Var(Y_{1,t-1}) + \frac{\alpha_1\alpha_2}{\beta_{11}\beta_{22}} Cov(Y_{1,t-1}, Y_{2,t-1}) \\
&+ \frac{(1-\beta_{12})(1-\beta_{21})}{\beta_{12}\beta_{21}} Cov(Y_{2,t-1}, Y_{1,t-1}) + \frac{(1-\beta_{21}\alpha_2)}{\beta_{21}\beta_{22}} Var(Y_{2,t-1})
\end{aligned}
$$

The first and second moments of $Y_{2,t}$ can be derived in a similar way. □

Proposition 8.4. *If the eigen-values η_1, η_2 of the following matrix*

$$
A = \begin{bmatrix} \frac{\alpha_1}{\beta_{11}} & \frac{1-\beta_{21}}{\beta_{21}} \\ \frac{1-\beta_{12}}{\beta_{12}} & \frac{\alpha_2}{\beta_{22}} \end{bmatrix} \tag{8.27}
$$

lie in the interval $[-1, 1]$, then the bivariate population X_t, Y_t will become extinct eventually.

Proof. The first moment can be expressed in a matrix form

$$
\mathbb{E}[\mathbf{Y}_t] = A\mathbb{E}[\mathbf{Y}_{t-1}] = A^t\mathbb{E}[\mathbf{Y}_0] \tag{8.28}
$$

The t-th power of a matrix here is defined as t times matrix multiplication. By eigen-decomposition, power of a matrix can be expressed as

$$
A^t = Q\,\mathrm{diag}(\{\eta_1^t, \eta_2^t\})Q^{-1}, \tag{8.29}
$$

where $Q = (\nu_1, \nu_2)$ is eigen vector matrix with ν_1, ν_2 as eigen vectors for η_1, η_2. Now, it is clear that $\mathbb{E}[\mathbf{Y}_t]$ is decreasing in t when $\eta_1, \eta_2 \in [-1, 1]$. □

8.4 Weak Convergence to continuous Birth and Death process

In this section, we will construct two continuous processes from the above proposed INAR models. These processes, under a certain parametrization, will converge weakly to the aforementioned continuous birth and death processes when the length of sub-interval goes to 0.

8.4.1 Construction of continuous processes

Since the continuous birth and death processes are clearly semimartingale defined in non-negative state spaces, to apply limit theorem of locally bounded semimartingales, we need to construct 'continuous' processes on a dense subsets of $\mathbb{R}_+$ (will take $t \in [0,1]$ for convenience) and compute their characteristic triplets from the discrete INAR models. Finally, when everything is set up nicely, we can apply weak convergence of semimartingale theorem to prove the result. The construction mainly follows from Jacod and Shiryaev (2013), Chapter II, section 3.

Starting with a discrete basis $\mathcal{B} = (\mathbf{\Omega}, \mathbf{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbf{P})$, assume that he INAR models X_n and $\mathbf{Y}_n$ defined above are adapted to this discrete stochastic basis and so as the increment processes

$$\begin{aligned} U_k &= X_k - X_{k-1}, \quad U_0 = X_0 \\ \mathbf{V}_k &= \mathbf{Y}_k - \mathbf{Y}_{k-1}, \quad \mathbf{V}_0 = \mathbf{Y}_0, \quad k = 0, 1, 2, \dots \end{aligned} \tag{8.30}$$

then we can construct 'continuous' processes via time change.

Definition 8.3. *Given a fixed time interval $[0,1]$, one can define a equal-length grid with size n such that each subinterval with length $\Delta = \frac{1}{n}$. The following the processes:*

$$Z_t^{(n)} = \sum_{k=0}^{\sigma_t} U_k, \quad \mathbf{M}_t^{(n)} = \sum_{k=0}^{\sigma_t} \mathbf{V}_t, \tag{8.31}$$

where $\sigma_t = \lfloor tn \rfloor$, are adapted to the continuous-time basis $\tilde{\mathcal{B}} = (\mathbf{\Omega}, \mathbf{F}, G = (\mathfrak{g}_t)_{t\geqslant 0}, \mathbf{P})$. The parameters setting for $Z_t^{(n)}$ are

$$\alpha = \frac{(\lambda-\mu)e^{(\lambda-\mu)\Delta}}{\lambda e^{(\lambda-\mu)\Delta} - \mu}, \quad p = \frac{\lambda-\mu}{\lambda e^{(\lambda-\mu)\Delta} - \mu}. \tag{8.32}$$

The parameters setting for $M_t^{(n)}$ are

$$\begin{aligned} \alpha_1 &= \frac{(\lambda_{11}-\mu_1)\omega_1(\Delta)}{\lambda_{11}\omega_1(\Delta)-\mu_1}, \quad \alpha_2 = \frac{(\lambda_{22}-\mu_2)\omega_2(\Delta)}{\lambda_{22}\omega_2(\Delta)-\mu_2} \\ \beta_{11} &= \frac{\lambda_{11}-\mu_1}{\lambda_{11}\omega_1(\Delta)-\mu_1}, \quad \beta_{22} = \frac{\lambda_{22}-\mu_2}{\lambda_{22}\omega_2(\Delta)-\mu_2} \\ \beta_{21} &= \left(1 + C_{\beta_1}\left(e^{u_1\Delta} - e^{u_2\Delta}\right)\right)^{-1}, \quad \beta_{12} = \left(1 + C_{\beta_2}\left(e^{u_1\Delta} - e^{u_2\Delta}\right)\right)^{-1}, \end{aligned} \tag{8.33}$$

where

$$\begin{aligned}
\omega_1(\Delta) &= C_\alpha e^{u_1\Delta} + (1-C_\alpha)e^{u_2\Delta}, \quad \omega_2(\Delta) = (1-C_\alpha)e^{u_1\Delta} + C_\alpha e^{u_2\Delta} \\
C_\alpha &= \frac{\lambda_{12}c}{2\lambda_{12}c+\mu_1-\mu_2},\ C_{\beta_1} = \frac{\lambda_{21}}{2\lambda_{12}c+\kappa_1-\kappa_2},\ C_{\beta_2} = \frac{\lambda_{12}}{2\lambda_{12}c+\kappa_1-\kappa_2} \\
u_1 &= \lambda_{12}c-\kappa_2, \quad u_2 = -(\lambda_{12}c+\kappa_1), \quad \kappa_i = \mu_i - \lambda_{ii},\ i=1,2 \\
c &= \frac{\kappa_2-\kappa_1+\sqrt{(\kappa_1-\kappa_2)^2+4\lambda_{21}\lambda_{12}}}{2\lambda_{12}}.
\end{aligned}$$

It is straightforward to derive the parameter setting for univariate case since we only need to match the parameter via probability generating function between $Z_t^{(n)}$ and Z_t. However, in the other case where the closed form probability generating function for $\mathbf{M}_t$ is not available, we need to seek other ways to set up α_i and $\beta_{i,j}$ in terms of λ and μ. The direct approach would be to match the first and second order moments to see whether it works. It is clear that we can match moment equations (8.26) to (8.16) and find out the mapping of β_{12}, β_{21} in terms of $\lambda_{i,j},\ \mu_i,\ i,j \in \{1,2\}$. Unfortunately, only the ratio α_i/β_{ii} is known. Nevertheless, the parameter setting in univariate case shows us the way to distribute the ratio α/p to α and p. Then α_i, β_{ii} can be set up in a similar way.

Proposition 8.5. *With the above parameters setting and any non-negative integer m, the transition probabilities for $Z_t^{(n)}$ conditional on $Z_{t-\Delta}^{(n)} = k$ are*

$$\begin{aligned}
\Pr(Z_t^{(n)} = k+m | Z_{t-\Delta}^{(n)} = k) &= \binom{k+m-1}{k-1}(\lambda\Delta)^m + o(\Delta^m) \\
\Pr(Z_t^{(n)} = k-m | Z_{t-\Delta}^{(n)} = k) &= \binom{k}{k-m}(\mu\Delta)^m + o(\Delta^m)
\end{aligned} \tag{8.34}$$

The above probabilities can be simplified as,

$$\begin{aligned}
\Pr\left(Z_t^{(n)} = k+1 | Z_{t-\Delta}^{(n)} = k\right) &= \lambda k\Delta + o(\Delta) \\
\Pr\left(Z_t^{(n)} = k-1 | Z_{t-\Delta}^{(n)} = k\right) &= \mu k\Delta + o(\Delta) \\
\Pr\left(|Z_t^{(n)} - k| \geqslant 2 | Z_{t-\Delta}^{(n)} = k\right) &= o(\Delta)
\end{aligned} \tag{8.35}$$

On the other hand, the transition probabilities for $\mathbf{M}_t^{(n)}$ conditional on $\mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k} =$

(k_1, k_2) *given by*

$$\begin{aligned}
&\Pr(M_{i,t}^{(n)} = k_i + m | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) \\
&= \sum_{j=k_i}^{k_i+m} \binom{j-1}{k_i-1} \binom{k_1+k_2+m-j-1}{k_{i'}-1} (\lambda_{ii}\Delta)^{j-k_i} (\lambda_{i',i}\Delta)^{k_i+m-j} + o(\Delta^m)
\end{aligned} \tag{8.36}$$

$$\Pr(M_{i,t}^{(n)} = k_i - m | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) = \binom{k_i}{k_i-m} (\mu_i\Delta)^m + o(\Delta^m),$$

where $i \in \{1, 2\}$ *and* $i' = 3 - i$. *Due to the conditional independence of bivariate INAR models, the joint transition probabilities for* $\mathbf{M}_t^{(n)}$ *conditional on* $\mathbf{M}_{t-\Delta}^{(n)}$ *are*

$$\begin{aligned}
&\Pr(M_{1,t}^{(n)} = k_1 \pm m_1, M_{2,t}^{(n)} = k_2 \pm m_2 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) \\
&= \Pr(M_{1,t}^{(n)} = k_1 \pm m_1 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) \Pr(M_{2,t}^{(n)} = k_2 + m_2 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k})
\end{aligned} \tag{8.37}$$

Similarly, the above probabilities can be simplified as

$$\begin{aligned}
&\Pr(M_{i,t}^{(n)} = k_i + 1 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) = \lambda_{ii} k_1 \Delta + \lambda_{i',i} k_2 \Delta + o(\Delta) \\
&\Pr(M_{i,t}^{(n)} = k_i - 1 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}) = \mu_i k_i \Delta + o(\Delta) \\
&\Pr\left(|M_{i,t}^{(n)} - k_i| \geqslant 2 | \mathbf{M}_{t-\Delta}^{(n)} = \mathbf{k}\right) = o(\Delta)
\end{aligned} \tag{8.38}$$

Proof. See Appendix A.8.C □

It is obvious that the above transition probabilities have exactly the same form as continuous counterparts when $m = 1$. Consequently, the Lévy measures of $Z_t^{(n)}$ and $\mathbf{M}_t^{(n)}$ have similar structure to their continuous counterparts.

Proposition 8.6. *The continuous processes* $Z_t^{(n)}$ *and* $\mathbf{M}_t^{(n)}$ *defined above are semi-*

martingales with following characteristics triplets.

$$Ch(Z_t^{(n)}) = \begin{cases} B_t = 0 \\ C_t = 0 \\ \nu([0,t] \times g) = \sum_{k=1}^{\sigma_t} (g(1)\lambda + g(-1)\mu) X_{k-1}\Delta + O(\Delta) \end{cases}$$

$$Ch(\mathbf{M}_t^{(n)}) = \begin{cases} \mathbf{B}_t = 0 \\ \mathbf{C}_t = 0 \\ \nu([0,t] \times g) = \sum_{k=1}^{\sigma_t} \left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 \right) \mathbf{Y}_{k-1}\Delta \\ \qquad\qquad + \left(g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(0,-1)\tilde{\boldsymbol{\mu}}_2 \right) \mathbf{Y}_{k-1}\Delta \\ \qquad\qquad + O(\Delta), \end{cases} \tag{8.39}$$

where the g is a continuous, non-negative, bounded Borel function vanishing near 0 and $\mathbf{M}_t^{(n)}$ respectively, the truncation function is $h = |x|\mathbf{1}_{\{|x|<1\}}$ and

$$\tilde{\boldsymbol{\lambda}}_1 = (\lambda_{11}, \lambda_{21}), \quad \tilde{\boldsymbol{\lambda}}_2 = (\lambda_{21}, \lambda_{22}), \quad \tilde{\boldsymbol{\mu}}_1 = (\mu_1, 0), \quad \tilde{\boldsymbol{\mu}}_2 = (0, \mu_2)$$

Proof. See appendix A.8.D □

Theorem 8.1. *With the the definition and the parametrization above, and the initial distribution condition:*

$$Z_0^{(n)} = Z_0, \quad \mathbf{M}_0^{(n)} = \mathbf{M}_0, \tag{8.40}$$

the processes $Z_t^{(n)}$ and $\mathbf{M}_t^{(n)}$ converge weakly to the continuous birth and death processes Z_t and $\mathbf{M}_t$.

$$\begin{aligned} \lim_{n\to\infty} Z_t^{(n)} &\xrightarrow{w} Z_t \\ \lim_{n\to\infty} \mathbf{M}_t^{(n)} &\xrightarrow{w} \mathbf{M}_t, \end{aligned} \tag{8.41}$$

when the size of subinterval Δ goes to 0 or equivalently, $n \to \infty$.

Proof. Here we simply apply Theorem 3.39 from Jacod and Shiryaev (2013), chapter IX ,section 3, the limit theorem of semimartingales for the locally bounded case.

i The local strong Majorization Hypothesis: For both cases Z_t and $\mathbf{M}_t$, the

first two terms of the characteristic triplets are 0 and stochastic integrals with respect to the function is clearly finite on $[0,1]$

ii Local Conditions on big jumps: For both cases Z_t and $\mathbf{M}_t$, there is no jump with absolute size greater than 1.

iii The local uniqueness: for every choices of initial distributions for Z_0 and $\mathbf{M}_0$, their Lévy measures are uniquely characterized by their (joint) probability distribution functions.

iv Continuity Condition, the characteristic triplets $B_t(\omega), C_t(\omega), \nu(\omega; dt, dx)$ of Z_t and $\mathbf{M}_t$ are continuous with respect to ω.

v Weak convergence of initial distribution. This is stated at the beginning of this theorem.

vi Convergence of characteristic triplet of discrete processes to that of their continuous counterparts. This can be proved by showing the uniform convergence of Lévy measures. For every $a > 0$, define a stopping time for the population process:

$$S_a(X) = \inf\{t : |X_t| > a, \text{ or } |X_{t-}| > a\} \tag{8.42}$$

For the univariate case, the stochastic integral with respect to $g * v$ for any Borel function g is given by

$$\begin{aligned}
&(g * \nu_{t\wedge S_a}) \circ Z^{(n)} = g * \nu(Z^{(n)}; [0, t \wedge S_a(Z^{(n)})], R) \\
=&\int_0^{t\wedge S_a(Z^{(n)})} \int_R g(x)(\lambda\delta_1(dx) + \mu\delta_{-1}(dx)) Z^{(n)}_{s-} ds \\
=&\int_0^{t\wedge S_a(Z^{(n)})} (g(1)\lambda + g(-1)\mu) Z^{(n)}_{s-} ds \\
=&\sum_{k=1}^{\sigma_{t\wedge S_a(Z^{(n)})}} (g(1)\lambda + g(-1)\mu)\, Z^{(n)}_{k-1}\Delta \\
&+ (g(1)\lambda + g(-1)\mu)\, Z^{(n)}_{\sigma_{t\wedge S_a(Z^{(n)})}} \left(t \wedge S_a(Z^{(n)}) - \sigma_{t\wedge S_a(Z^{(n)})}\Delta\right)
\end{aligned} \tag{8.43}$$

and the absolute difference of two stochastic integrals is given by,

$$\begin{aligned}
&|g * \nu^n_{t\wedge S_a} - (g * \nu_{t\wedge S_a}) \circ Z^{(n)}| \\
=&\left|O(\Delta) + (g(1)\lambda + g(-1)\mu)\, Z_{\sigma_{t\wedge S_a(Z^{(n)})}} \left(t \wedge S_a(Z^{(n)}) - \sigma_{t\wedge S_a(Z^{(n)})}\Delta\right)\right| \\
\leqslant&O(\Delta) + |g(1)\lambda + g(-1)\mu| Z_{\sigma_{t\wedge S_a(Z^{(n)})}} \left(t \wedge S_a(Z^{(n)}) - \sigma_{t\wedge S_a(Z^{(n)})}\Delta\right)
\end{aligned} \tag{8.44}$$

It is clear that all the quantity inside $|..|$ are finite and for every $\xi > 0$, and then there exists a natural number N such that for $n > N$, we have

$$|g * \nu^n_{t\wedge S_a} - (g * \nu_{t\wedge S_a}) \circ Z^{(n)}| < \xi \tag{8.45}$$

and hence we have the uniform convergence for $g * \nu^n_{t\wedge S_a}$ to $(g * \nu_{t\wedge S_a}) \circ Z^{(n)}$. For the bivariate case, the stochastic integral $g * \nu$, where ν is the Lévy measure of M, for any Borel function g is given by

$$\begin{aligned}
&(g * \nu_{t\wedge S_a}) \circ \mathbf{M}^{(n)} = g * \nu(\mathbf{M}^{(n)}; [0, t \wedge S_a(\mathbf{M}^{(n)})], R) \\
=&\int_0^{t\wedge S_a(\mathbf{M}^{(n)})} \int_R g(x)(\tilde{\boldsymbol{\lambda}}_1 \delta_{(1,0)}(dx) + \tilde{\boldsymbol{\lambda}}_2 \delta_{(0,1)}(dx) \\
&+ \tilde{\boldsymbol{\mu}}_1 \delta_{(0,-1)}(dx) + \tilde{\boldsymbol{\mu}}_2 \delta_{(0,-1)}(dx)) \mathbf{M}^{(n)}_{s^-} ds \\
=&\int_0^{t\wedge S_a(\mathbf{M}^{(n)})} \left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,-1)\tilde{\boldsymbol{\mu}}_2 \right) \mathbf{M}^{(n)}_{s^-} ds \\
=&\sum_{k=1}^{t\wedge S_a(\mathbf{M}^{(n)})} \left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,-1)\tilde{\boldsymbol{\mu}}_2 \right) \mathbf{M}^{(n)}_{k-1} \Delta \\
&+ \left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,-1)\tilde{\boldsymbol{\mu}}_2 \right) \\
&\times \mathbf{M}^{(n)}_{\sigma_{t\wedge S_a(\mathbf{M}^{(n)})}} \left(t \wedge S_a(\mathbf{M}^{(n)}) - \sigma_{t\wedge S_a(\mathbf{M}^{(n)})} \right)
\end{aligned} \tag{8.46}$$

Then the absolute difference of two stochastic integrals is given by

$$\begin{aligned}
&|g * \nu_{t\wedge S_a(\mathbf{M}^{(n)})} - (g * \nu_{t\wedge S_a}) \circ \mathbf{M}^{(n)}| \\
\leqslant& O(\Delta) + \left| g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,-1)\tilde{\boldsymbol{\mu}}_2 \right| \\
&\times \mathbf{M}^{(n)}_{\sigma_{t\wedge S_a(\mathbf{M}^{(n)})}} \left(t \wedge S_a(\mathbf{M}^{(n)}) - \sigma_{t\wedge S_a(\mathbf{M}^{(n)})} \right)
\end{aligned} \tag{8.47}$$

Hence the uniform convergence holds using similar argument as in the univariate case. Finally, the $Z^{(n)}_t$, $M^{(n)}_t$ converge weakly to Z_t and M_t respectively. □

8.5 Simulation Study

In this section, we outline the simulation algorithm for bivariate birth and death processes. Then estimation method, properties of estimators are investigated in the simulation study.

8.5.1 Simulation of bivariate birth and death process

The simulation algorithm of bivariate birth and death process $\mathbf{M}_t$ can be derived straightforwardly according to its ODE (8.11). Given the current population $\mathbf{M}_t$, the waiting time that a event (birth or death in either population) will happen follows exponential distribution with rate

$$\rho_t = (\lambda_{11} + \lambda_{12} + \mu_1)M_{1,t} + (\lambda_{21} + \lambda_{22} + \mu_2)M_{2,t}$$

Then the probability that this event will happen in population $M_{1,t}$ is

$$p_1 = \frac{\lambda_{21}M_{2,t} + (\lambda_{11} + \mu_1)M_{1,t}}{\rho_t} \tag{8.48}$$

The probability that this event will happen in population $M_{2,t}$ would simply be $p_2 = 1 - p_1$. Suppose now an event happens in population $M_{1,t}$, the probability that there is a new individual would be

$$p_1^b = \frac{\lambda_{11}M_{1,t} + \lambda_{21}M_{2,t}}{\lambda_{21}M_{2,t} + (\lambda_{11} + \mu_1)M_{1,t}}, \tag{8.49}$$

and the probability that an individual dies is $p_1^d = 1 - p_1^b$. Likewise, if the event happens in the population $M_{2,t}$, the birth probability would be

$$p_2^b = \frac{\lambda_{12}M_{1,t} + \lambda_{22}M_{2,t}}{\lambda_{12}M_{1,t} + (\lambda_{22} + \mu_2)M_{2,t}} \tag{8.50}$$

and death probability $p_n^d = 1 - p_n^b$. Overall, the simulation algorithm is shown in the following algorithm 1.

On the other hand, the simulation procedure of bivariate INAR(1) model is straightforward because the distribution of $\mathbf{Y}_t$ are indicated by the operator $(\circ,\ *_1,\ *_2)$ given $\mathbf{Y}_{t-1}$.

Algorithm 1 Simulation of bivariate birth and death process with rates $\{\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}, \mu_1, \mu_2\}$, initial population $M_{1,0}, M_{2,0}$, a vector of cumulative time t_c where $t_c[1] = 0$, a counter i and terminal time T

1. Simulate a waiting time $t_w \sim Exp(\rho_t)$ and two independent uniform random variable $U_1, U_2 \sim U(0,1)$
2. if $U_1 \leqslant p_1$, and $U_2 \leqslant p_1^b$, $M_{1,t+t_w} = M_{1,t} + 1$ and $M_{2,t+t_w} = M_{2,t}$
3. if $U_1 \leqslant p_1$, and $U_2 > p_1^b$, $M_{1,t+t_w} = M_{1,t} - 1$ and $M_{2,t+t_w} = M_{2,t}$
4. if $U_1 > p_1$, and $U_2 \leqslant p_2^b$, $M_{2,t+t_w} = M_{2,t} + 1$ and $M_{1,t+t_w} = M_{1,t}$
5. if $U_1 > p_1$, and $U_2 > p_2^b$, $M_{2,t+t_w} = M_{2,t} - 1$ and $M_{1,t+t_w} = M_{1,t}$
6. Append a new element to t_c, $t_c[i+1] = t_c[i] + t_w$ and update counter $i = i + 1$
7. Repeat all the steps above until $t_c[i] > T$ or $M_{1,t_c[i]} = M_{2,t_c[i]} = 0$
8. Set $M_{1,T} = M_{1,t_c[i-1]}$ and $M_{2,T} = M_{2,t_c[i-1]}$ and return the trajectory $\mathbf{M}$ at each element of t_c

8.5.2 Statistical inference of Univariate and bivariate birth and death process

8.5.2.1 Quasi-MLE for univariate LBD

In the univariate case, parameters estimation and their asymptotic properties are available in Keiding (1975). Suppose now we have the full information of the sample path, the exact inter-arrival times for each birth and death events $\{\tau_i\}_{\{i=0,1,2,\dots\}}$ on the sampling interval $[0, T]$ where $\tau_0 = 0$, the maximum likelihood estimators for Z_t are

$$\hat{\lambda} = \frac{B_T}{X_T}, \quad \hat{\mu} = \frac{D_T}{X_T}, \quad X_T = \sum_{k=1}^{B_T+D_T} \tau_k Z_{\tau_{k-1}} + \left(T - \sum_{i=1}^{n} \tau_i\right) Z_T, \tag{8.51}$$

where B_T, D_T are total number of birth and death events respectively. The asymptotic properties are given by fixed T and large population

$$\lim_{Z_0 \to \infty} \left(\frac{Z_0(e^{(\lambda-\mu)T} - 1)}{\lambda - \mu}\right)^{\frac{1}{2}} \begin{pmatrix} \hat{\lambda} - \lambda \\ \hat{\mu} - \mu \end{pmatrix} \xrightarrow{D} \mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}\right) \tag{8.52}$$

In practice, one may not have exact information of inter-arrival time of the events. Instead, we have records for populations sampling over a fixed-length interval Δ such that $Z_0, Z_\Delta, Z_{2\Delta} \dots Z_{n\Delta}$ are available. Then to estimate the parameters λ, μ, one can numerically maximize the Quasi log-likelihood function from the proposed INAR(1) model $X_k = Z_{k\Delta},\ k = 0, 1, \dots, n$. The log likelihood function is given by,

$$\ell(\alpha, p) = \sum_{k=1}^{n} \log \Pr(X_{k-1}, X_k)$$

$$\Pr(X_{k-1}, X_k) = \begin{cases} 1, & X_{k-1} = X_k = 0 \\ (1-\alpha)^{X_{t-1}}, & X_k = 0 \\ \sum_{j=1}^{\min\{X_{k-1}, X_k\}} f_b(j; X_{k-1}, \alpha) f_{nb}(X_k - j; j, p), & X_{k-1} > 0 \ \& \ X_k > 0, \end{cases} \tag{8.53}$$

where f_b and f_{nb} are probability mass function of binomial and negative binomial random variables

$$f_b(x; n, \alpha) = \binom{n}{x} \alpha^x (1-\alpha)^{n-x} \quad f_{nb}(x; n, \beta) = \binom{n+x-1}{n-1} \beta^n (1-\beta)^x$$

The simulation is conducted as follow: we generate 1000 sample paths of Z_t using

Table 8.1: Parameter setting for univariate Case

parameters	λ	μ	Z_0	T
values	1.2	1	100	1

the parameters settings in Table 8.1. Since Z_t are continuous sample paths, we set up an equal-distance grid with sampling interval Δ. Then the equal-distance observations X_t are obtained by counting the total number of population up to each discrete time $(0, \Delta, 2\Delta, \dots, n\Delta)$ where $n = \frac{T}{\Delta}$. The log likelihood function is then maximized by 'optim' function with method = 'BFGS' in R programming. Finally, we can recover the rate estimates by inverting the parametrization in equation (8.32) such that

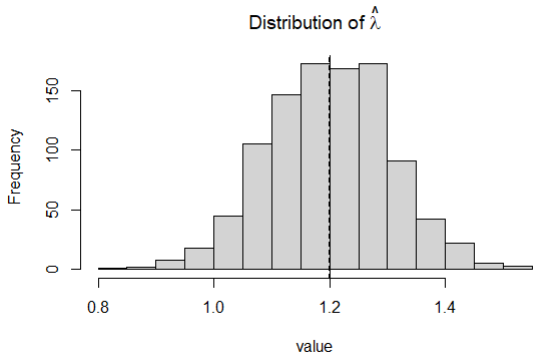
$$\tilde{\lambda} = \frac{\frac{1-\hat{p}}{\hat{p}} \log \frac{\hat{\alpha}}{\hat{p}}}{\frac{\hat{\alpha}}{\hat{p}} - 1}, \quad \tilde{\mu} = \tilde{\lambda} - \frac{1}{\Delta} \log \frac{\hat{\alpha}}{\hat{p}} \tag{8.54}$$

In the following, we will first explore how the size of Δ would affect properties estimators, i.e. bias and mean square error (MSE), and how much more computational time we need compared to true MLE method. Four different size of sampling inter-

vals $\Delta = \{0.1, 0.05, 0.025, 0.01\}$ is chosen and the results are presented in table 8.2. The theoretical row shows the biased and MSE computed through equation (8.52). There is no surprise that the True MLE method from equation (8.51) performs the best, with lowest MSE and computational time. The Quasi-MLE method by constructing INAR model, on the other hand, becomes better as we decreasing the size of sampling interval Δ but it still performs no better than the true MLE method and require much more computational time. The empirical distribution of these estimators are illustrated in figure 8.1 and since the general shape of distribution of $\tilde{\lambda}$ and $\tilde{\mu}$ has little difference, we will only show the distribution of $\tilde{\lambda}$. It is clear that only the case $\Delta = 0.01$ has satisfactory normal shape compared to all other cases.

Table 8.2: Properties of different maximum likelihood estimators. The time column is the total time of estimating 1000 sample paths.

	Bias λ	MSE λ	Bias μ	MSE μ	time (s)
Theoretical	0	0.000126	0	0.000116	-
MLE	-0.000701	0.000124	0.015898	0.000396	0.09
$\Delta = 0.1$	-0.133845	0.075696	-0.116943	0.072754	49
$\Delta = 0.05$	-0.086435	0.023935	-0.069513	0.021836	98.6
$\Delta = 0.025$	-0.051234	0.007711	-0.034323	0.006557	201.4
$\Delta = 0.01$	-0.014178	0.001378	0.002721	0.001159	522.2



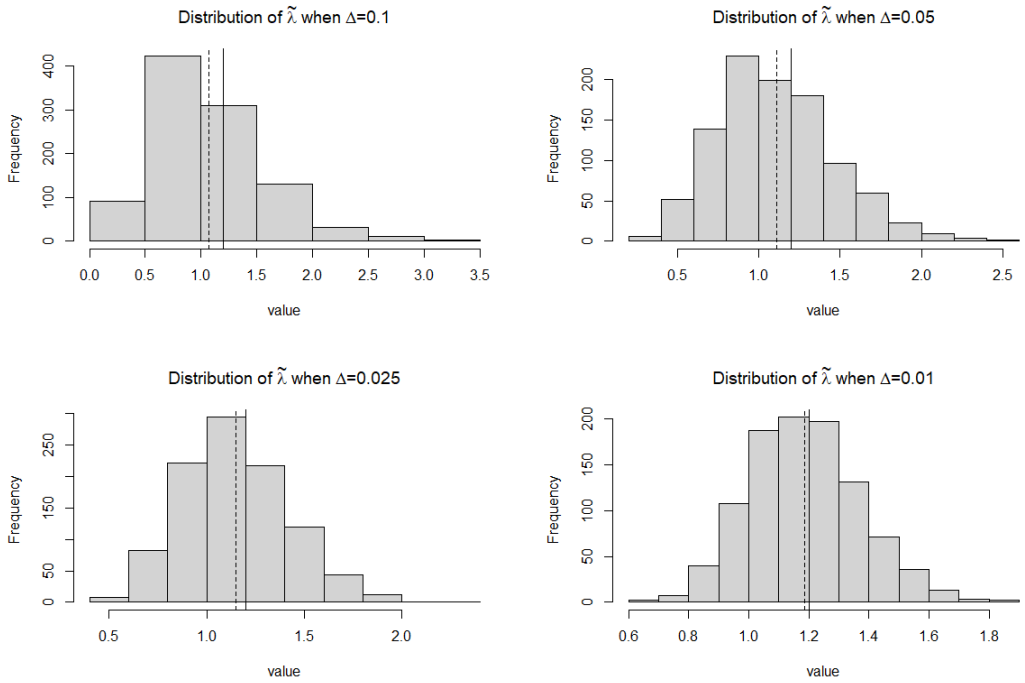


Figure 8.1: The empirical distribution of estimated parameters. The top panel is the MLE from 8.51 and the rest of plots are MLE from INAR model. The solid lines are the true values of the parameters listed in table 8.1 and the dash lines stand for empirical means.

To achieve asymptotic normality for Quasi-MLE method from INAR model, one need not only large initial population, but also a small sampling interval Δ. In the following simulation, we would fix the sampling interval $\Delta = 0.01$ and investigate how the size of initial population would affect the asymptotic distribution of estimators and the computational time for estimation procedure. To explore the effect of Z_0 for asymptotic distribution, we choose $Z_0 \in \{5, 10, 30, 50\}$ and it seems from Figure 8.2 that to ensure asymptotic normality for both estimators, one need at least $Z_0 = 30$, which is a large sample size in statistical sense.

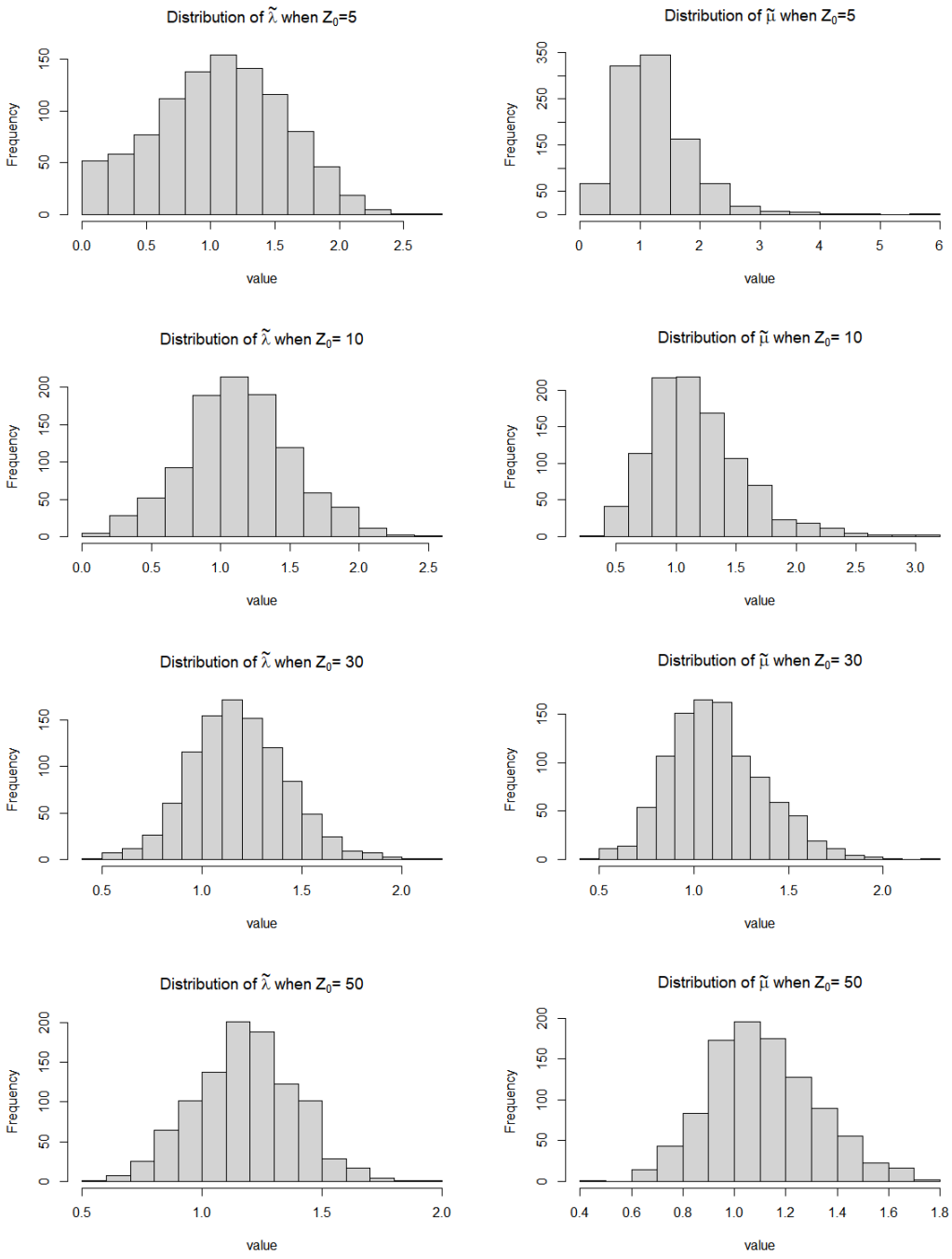


Figure 8.2: Asymptotic distribution of $\tilde{\lambda}$, $\tilde{\mu}$ with different Z_0

The computational time with respect to $Z_0 \in \{10, 50, 100, 150, \ldots, 500\}$ clearly shows a linear trend in 8.3. This is reasonable as the number of summation involved in equation (8.53) increases linearly with respect to Z_0

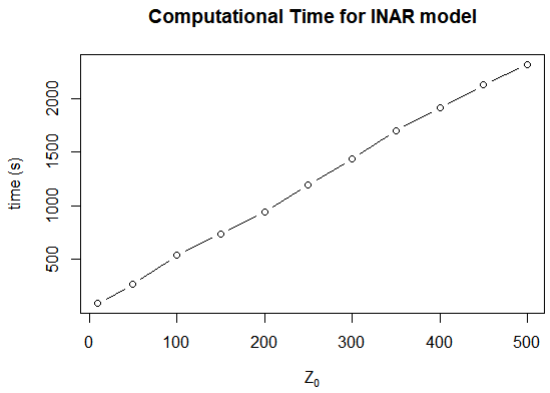


Figure 8.3: The computational time for INAR models of 1000 sample paths

In summary, the Quasi-MLE method constructed from INAR model can reach moderate level of estimation accuracy and asymptotic normality with large initial population $Z_0 \geqslant 30$ and small sampling interval $\Delta \leqslant 0.01$. However, it would require much more computational time than the true MLE method. This method should only be used in the case where we have no information on inter-arrival time of birth and death events.

8.5.2.2 Quasi-MLE for bivariate LBD

Since the bivariate INAR(1) model is a bivariate Markov Chain, the log likelihood function can be written as the sum of logarithm of transition probabilities. Denote $\Theta = \{\alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$ as the parameter space of bivariate INAR(1) model, then the likelihood function can be written as

$$\begin{aligned}
\ell(\Theta) &= \sum_{t=1}^{n} \log \Pr(X_t, Y_t | X_{t-1}, Y_{t-1}) \\
&= \sum_{t=1}^{n} \left(\log \Pr(X_t | X_{t-1}, Y_{t-1}) + \log \Pr(Y_t | X_{t-1}, Y_{t-1}) \right) \\
&= \ell_x(\Theta_x) + \ell_y(\Theta_y),
\end{aligned} \tag{8.55}$$

where $\Theta_x = \{\alpha_1, \beta_{11}, \beta_{21}\}$ and $\Theta_y = \{\alpha_2, \beta_{12}, \beta_{22}\}$. Because X_t and Y_t are independent of each other given the last state (X_{t-1}, Y_{t-t}), the likelihood function can be separated into two parts, ℓ_x and ℓ_y respectively. Then transition probability for X_t

is given by

$$\Pr(X_t = z_1 | X_{t-1} = x, Y_{t-1} = y)$$

$$= \begin{cases} 1 & z_1 = x = y = 0 \\ (1-\alpha_1)^x \beta_{21}^y & z_1 = 0 \\ f_{nb}(z_1; y, \beta_{21}) & x = 0 \ \& \ y > 0 \\ \sum_{i=1}^{\min\{x, z_1\}} f_b(i; x, \alpha_1) f_{nb}(z_1 - i; i, \beta_{11}) & x > 0 \ \& \ y = 0 \\ \sum_{j=1}^{z_1} \sum_{i=1}^{\min\{x,j\}} f_b(i; x, \alpha_1) f_{nb}(z_1 - i; i, \beta_{11}) f_{nb}(z_1 - j; y, \beta_{21}) & \\ \quad + (1-\alpha_1)^x f_{nb}(z_1; y, \beta_{21}) & x > 0 \ \& \ y > 0 \end{cases}$$

The one for Y_t is

$$\Pr(Y_t = z_2 | X_{t-1} = x, Y_{t-1} = y)$$

$$= \begin{cases} 1 & z_2 = x = y = 0 \\ (1-\alpha_2)^y \beta_{12}^x & z_2 = 0 \\ f_{nb}(z_2; x, \beta_{12}) & x > 0 \ \& \ y = 0 \\ \sum_{i=1}^{\min\{y, z_2\}} f_b(i; y, \alpha_2) f_{nb}(z_2 - i; i, \beta_{22}) & x = 0 \ \& \ y > 0 \\ \sum_{j=1}^{z_2} \sum_{i=1}^{\min\{y,j\}} f_b(i; y, \alpha_2) f_{nb}(z_2 - i; i, \beta_{22}) f_{nb}(z_2 - j; y, \beta_{12}) & \\ \quad + (1-\alpha_2)^y f_{nb}(z_2; x, \beta_{12}) & x > 0 \ \& \ y > 0 \end{cases}$$

One can then numerically maximize the log likelihood function ℓ_x, ℓ_y given the random samples $\{(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)\}$. From the estimated parameters $\hat{\Theta}$, we can solve the following system of equations to get the estimates $\Theta_{bd} = \{\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}, \mu_1, \mu_2\}$ for bivariate birth and death process.

$$\begin{cases} \alpha_1(\Theta_{bd}, \Delta) - \hat{\alpha}_1 = 0 \\ \alpha_2(\Theta_{bd}, \Delta) - \hat{\alpha}_2 = 0 \\ \beta_{11}(\Theta_{bd}, \Delta) - \hat{\beta}_{11} = 0 \\ \beta_{12}(\Theta_{bd}, \Delta) - \hat{\beta}_{12} = 0 \\ \beta_{21}(\Theta_{bd}, \Delta) - \hat{\beta}_{21} = 0 \\ \beta_{22}(\Theta_{bd}, \Delta) - \hat{\beta}_{22} = 0, \end{cases} \tag{8.56}$$

where the parametrization function $.(\Theta_{bd}, \Delta)$ are given in equation 8.33 and Δ is chosen based on the interpretation of birth and death rates. For example, when the random samples are collected on daily basis over a year $t = 1$, one can define $\Delta = t/365$. Then these parameters Θ_{bd} are interpreted on an annual scale.

Table 8.3: Parameter setting for simulation

parameters	λ_{11}	λ_{12}	λ_{21}	λ_{22}	μ_1	μ_2	T	$\mathbf{M}_0$
values	0.3	1.2	1.3	0.4	1.1	1.2	1	(40,50)

In the following, we will simulate the $r_2 = 1000$ sample paths of $\mathbf{M}_t$ based on the pre-specific parameters in table 8.3. Then equal-distance gird with sampling interval Δ is set up and random samples $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ are obtained, like the way mentioned in the univariate case. Then the likelihood functions ℓ_x, ℓ_y are maximized by 'optim' in **R** with method being specified as 'BFGS' and the maximum likelihood estimators $\hat{\Theta}$ are obtained. Finally, we can obtain the estimators $\hat{\Theta}_{bd}$ by numerically solving the system of equations (8.56) via a root-finding algorithm (e.g. Newton-Raphson method). Referring to the estimation results in univariate case, we focus on the choices of $\Delta \in \{0.02, 0.01, 0.005\}$ as well as large initial population $(40, 50)$, and hopefully we can obtain asymptotic normality for each estimator. The empirical distribution of these estimators Θ_{bd} are illustrated in Figure 8.4 and their properties are summarized in Table 8.4.

Table 8.4: Properties of different maximum likelihood estimators. The time column is the total time of estimating 1 sample paths.

Bias	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$	time (s)
$\Delta = 0.02$	-0.031393	0.069314	0.031804	-0.072310	0.025835	0.012040	25.46
$\Delta = 0.01$	-0.026080	0.061711	0.019728	-0.064421	0.018005	0.012375	46.7
$\Delta = 0.005$	0.002760	0.044764	-0.004686	-0.049816	0.019450	0.010779	82.32
MSE	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$	
$\Delta = 0.02$	0.266179	0.401756	0.301175	0.296895	0.104222	0.104958	
$\Delta = 0.01$	0.249248	0.359153	0.263573	0.270960	0.055740	0.058685	
$\Delta = 0.005$	0.266925	0.351415	0.268431	0.286881	0.037742	0.036170	

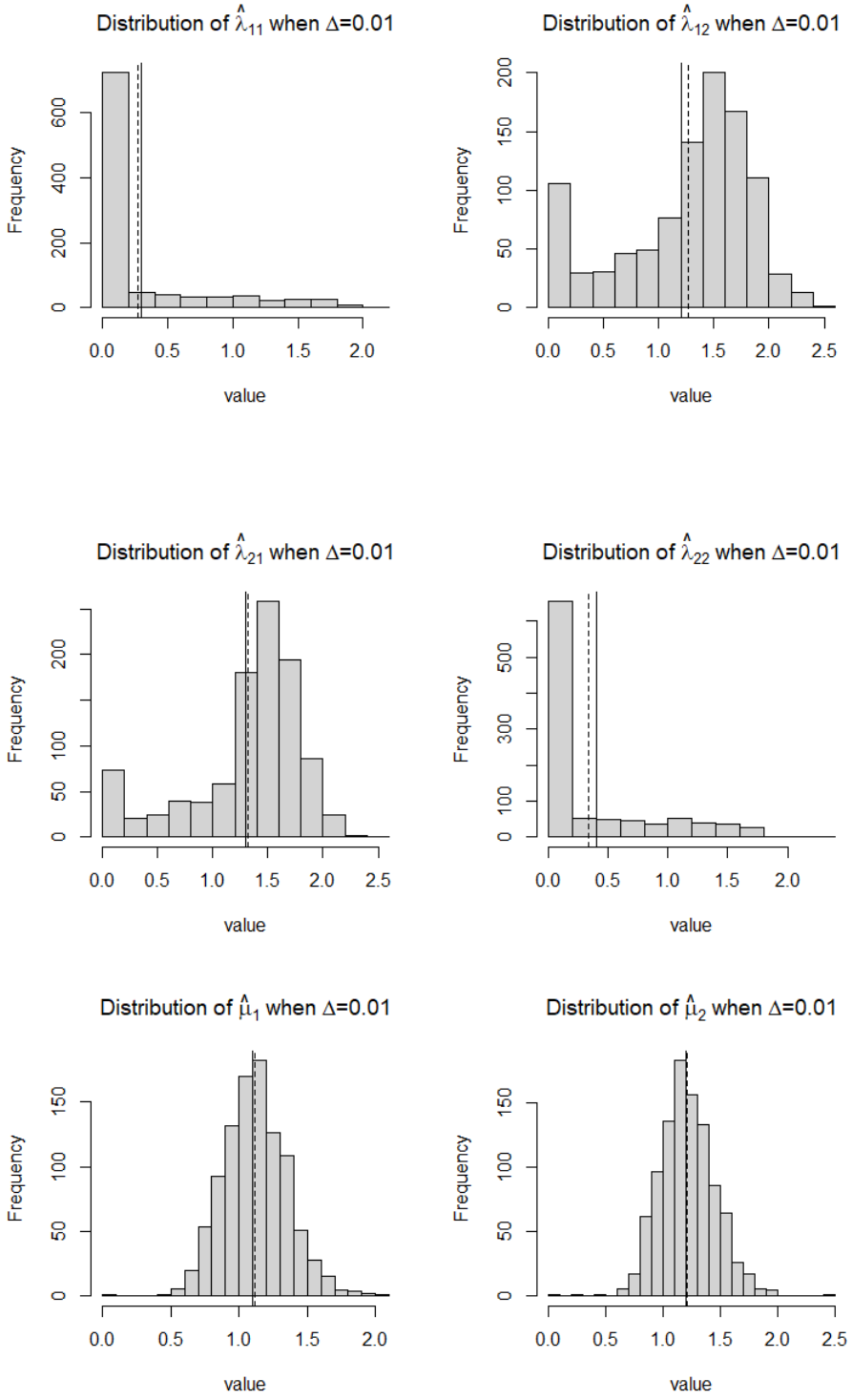


Figure 8.4: Empirical distribution estimators from bivariate INAR model. The solid lines are the true values of the parameters listed in table 8.3 and the dash lines stand for empirical means.

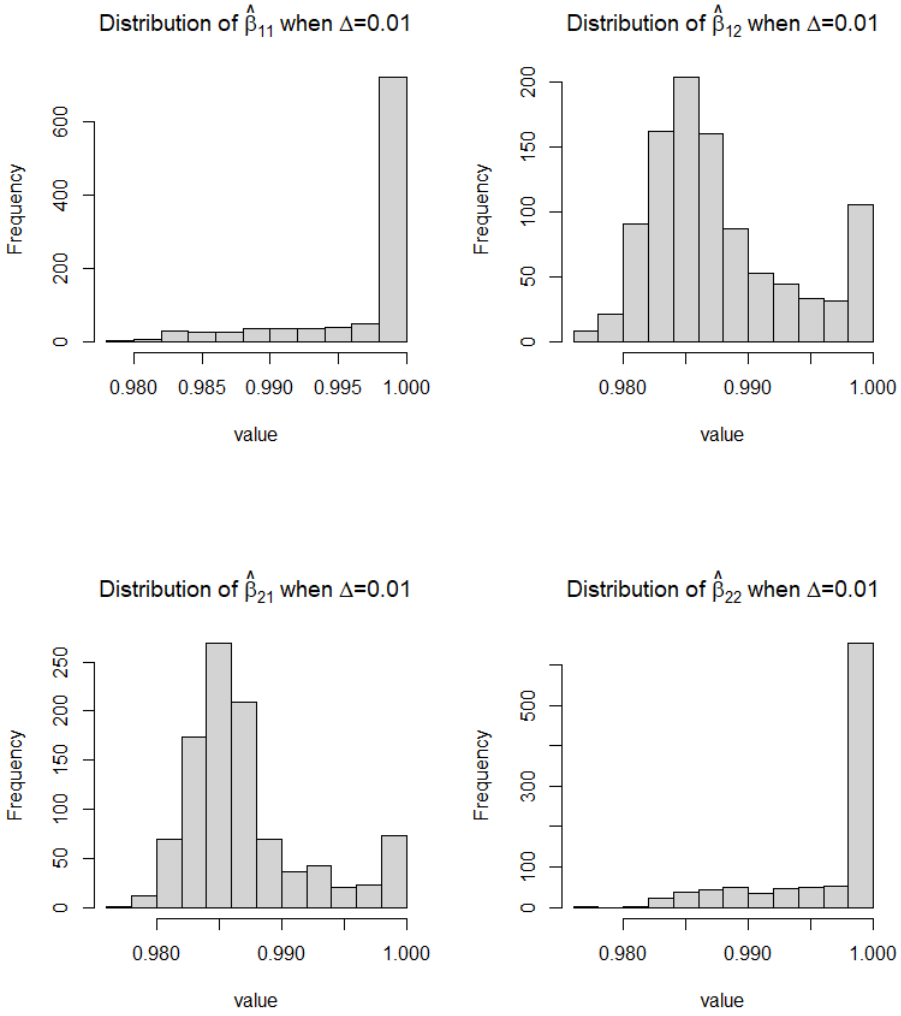


Figure 8.5: Empirical distribution estimators from bivariate INAR model.

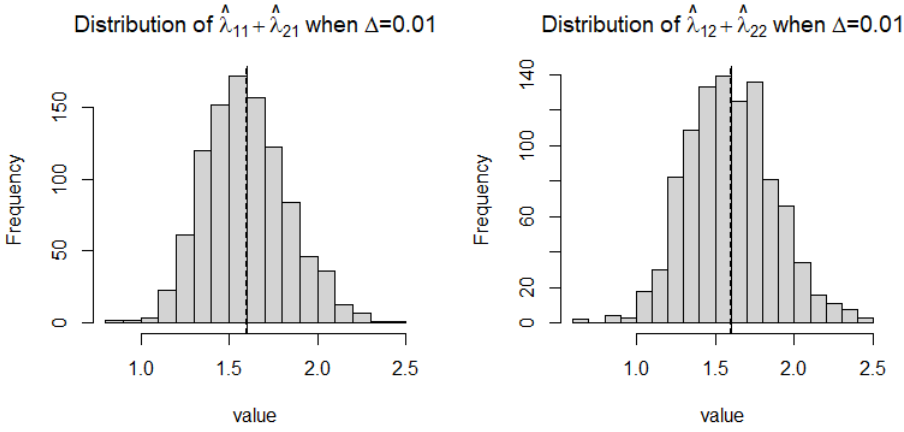


Figure 8.6: Empirical distribution for total birth rates. The solid lines are the true values of the parameters listed in table 8.3 and the dash lines stand for empirical means.

The bias and MSE of most estimators are decreasing with respect to Δ as expected.

However, the MSE of birth rates are much larger than the estimators of death rates. Except the estimators for death rates, all other estimators for birth rate are skewed to different directions and clearly non-normal distributed. This may caused by some of non-normal estimators for proposed INAR model illustrated in Figure 8.5. In the classical setting where the innovation term is included, one need stationary condition to ensure asymptotic normality for all estimators of parameters, see Bu et al. (2008). And in our case, INAR model itself is not stationary and hence some of the estimate can be skewed.

Notice that the pair of birth rates that contributed to the same population, $(\lambda_{11}, \lambda_{21})$ and $(\lambda_{12}, \lambda_{22})$ are skewed in opposite directions. It is then worthwhile to see whether the sum of these pair estimators has desired asymptotic properties and the results in Figure 8.6 confirms our conjecture. Combining the simulation procedure of bivariate birth and death processes, Quasi-MLE method may not be able to distinguish the pair of birth rates contributed to the same population. Instead, it would provide good estimators for the scale of total birth rates $\bar{\lambda}_1 = \hat{\lambda}_{11} r_m + \hat{\lambda}_{21}(1 - r_m)$ and $\bar{\lambda}_2 = \hat{\lambda}_{12} r_m + \hat{\lambda}_{22}(1 - r_m)$ where $r_m = \frac{\mathbb{E}[M_{1,t}]}{\mathbb{E}[M_{1,t}+M_{2,t}]}$. Furthermore, according to the proof A.8.A, the relationship between first moment of two population is given by

$$\mathbb{E}[M_{1,t}] = c\mathbb{E}[M_{2,t}] + (M_{1,0} - cM_{2,0})e^{-(\lambda_{12}c-\kappa_2)t}. \tag{8.57}$$

As long as the whole process is not extinct with probability one, i.e. $\kappa_1\kappa_2 < \lambda_{12}\lambda_{21}$, the exponential power $(\lambda_{12}c - \kappa_2)$ will always be positive and hence $\mathbb{E}[M_{1,t}] \approx c\mathbb{E}[M_{2,t}]$ when t is large. In other words, the ratio

$$r_m = \frac{\mathbb{E}[M_{1,t}]}{\mathbb{E}[M_{1,t} + M_{2,t}]} \to \frac{c}{1+c}, \tag{8.58}$$

becomes a constant eventually. For the parameter setting in Table 8.3, $c = 1.040833$, $r_m \approx \frac{1}{2}$ and hence $\hat{\lambda}_{11} + \hat{\lambda}_{21}$ serves as an estimator for the total birth rate of $M_{1,t}$. In practice, the c is unknown as true parameters need to be estimated. Then we can use the values at the end of sampling period to approximate r_m, i.e.

$$r_m \approx \frac{M_{1,T}}{M_{1,T} + M_{2,T}} \tag{8.59}$$

Table 8.5: Properties for total birth rates estimators

	bias $\bar{\lambda}_1$	MSE$\bar{\lambda}_1$	$Bias\bar{\lambda}_2$	MSE $\bar{\lambda}_2$
$\Delta = 0.02$	0.012623	0.024232	-0.015039	0.028914
$\Delta = 0.01$	0.008386	0.013068	-0.014407	0.017563
$\Delta = 0.005$	0.009380	0.009286	-0.015613	0.011202

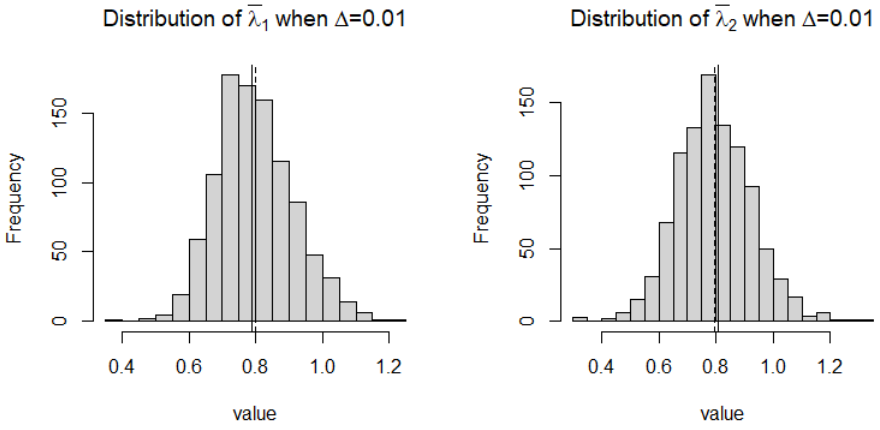


Figure 8.7: Empirical distribution for total birth rates. The solid lines are the true values of the parameters listed in table 8.3 and the dash lines stand for empirical means.

The properties of $\bar{\lambda}_1, \bar{\lambda}_2$ and their empirical distribution are shown in Table 8.5 and Figure 8.7. These new estimators benefits from nice properties, low bias and MSE and they decreases as Δ decreases. Most importantly, they are not skewed anymore and asymptotic normal.

Let us try another parameter setting in Table 8.6 to verify this conjecture. Same simulation and estimation process as previous case and the results are shown in Table 8.7, 8.8 and Figure 8.8. This time the constant c is 0.576306 and $r_m = 0.365605$. Similar to the last setting, the estimators for all birth rates are skewed and some of them have large bias and MSE. The estimators for total birth rate, on the other hand, are of low bias and MSE and they are again asymptotic normal.

Table 8.6: Parameter setting for simulation

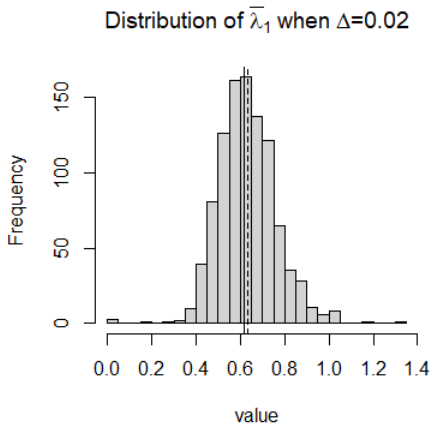
parameters	λ_{11}	λ_{12}	λ_{21}	λ_{22}	μ_1	μ_2	T	$\mathbf{M}_0$
values	0.3	0.5	0.8	1.5	1.1	1.2	1	(30,60)

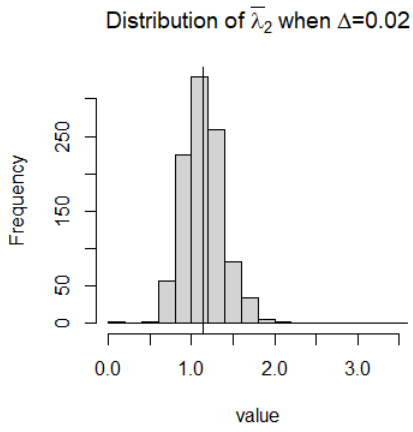
Table 8.7: Properties of different maximum likelihood estimators.

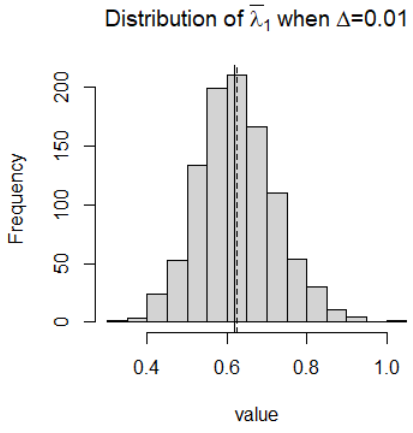
Bias	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\Delta = 0.02$	-0.091964	1.633439	0.064837	-0.891418	0.050226	0.008661
$\Delta = 0.01$	-0.092042	1.613498	0.056800	-0.863226	0.034508	0.02260
$\Delta = 0.005$	-0.063956	1.513447	0.036706	-0.818343	0.024054	0.016939
MSE	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\Delta = 0.02$	0.243023	4.669929	0.108730	1.315634	0.115385	0.120301
$\Delta = 0.01$	0.225980	4.535412	0.090212	1.259920	0.062935	0.06831
$\Delta = 0.005$	0.243969	4.186109	0.087121	1.177241	0.042955	0.036328

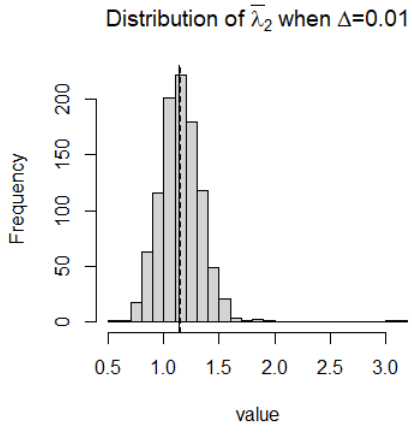
Table 8.8: Properties for total birth rates estimators

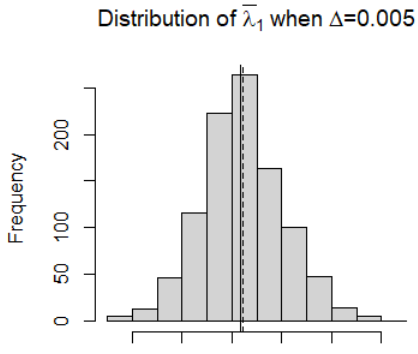
	Bias $\bar{\lambda}_1$	MSE $\bar{\lambda}_1$	Bias $\bar{\lambda}_2$	MSE $\bar{\lambda}_2$
$\Delta = 0.02$	0.013970	0.017193	0.007765	0.062907
$\Delta = 0.01$	0.008500	0.009268	0.017503	0.038616
$\Delta = 0.005$	0.004425	0.006884	0.012662	0.023027

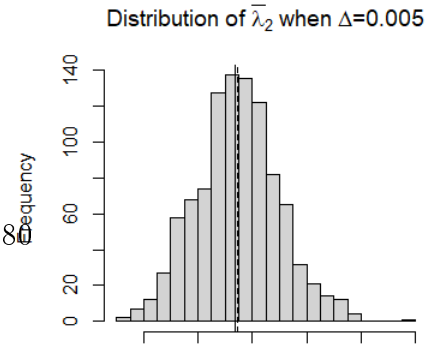












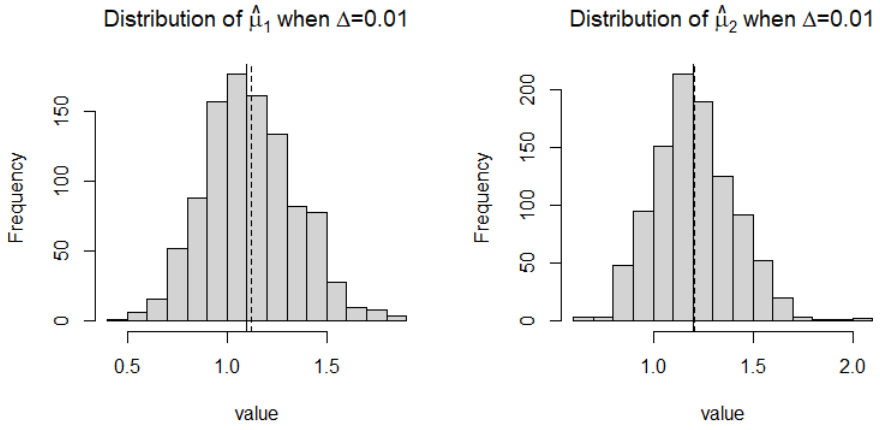
Let us finally try another parameter setting in Table 8.9 where the $\mathbf{M}_t$ is going to be extinct eventually. It means that the exponential function in equation (8.57) can no longer be omitted. The results are illustrated in Table 8.10 and Figure 8.9 and they look similar to the results of the first case. Nice properties for death rates' estimators but skewed and non-normal for birth rates' estimators.

Table 8.9: Parameter setting for simulation

parameters	λ_{11}	λ_{12}	λ_{21}	λ_{22}	μ_1	μ_2	T	$\mathbf{M}_0$
values	0.3	0.5	0.8	0.3	1.1	1.2	1	(30,60)

Table 8.10: Properties of different maximum likelihood estimators.

Bias	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\Delta = 0.02$	0.026802	0.071221	-0.020855	-0.064399	0.015807	0.007330
$\Delta = 0.01$	0.033741	0.063121	-0.024328	-0.056261	0.019261	0.008451
$\Delta = 0.005$	0.032433	0.061202	-0.022955	-0.046396	0.016784	0.011859
MSE	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\Delta = 0.02$	0.245305	0.184363	0.165970	0.091334	0.088017	0.059102
$\Delta = 0.01$	0.226033	0.167176	0.150903	0.081478	0.052522	0.039980
$\Delta = 0.005$	0.211617	0.155803	0.139283	0.078867	0.042573	0.035270



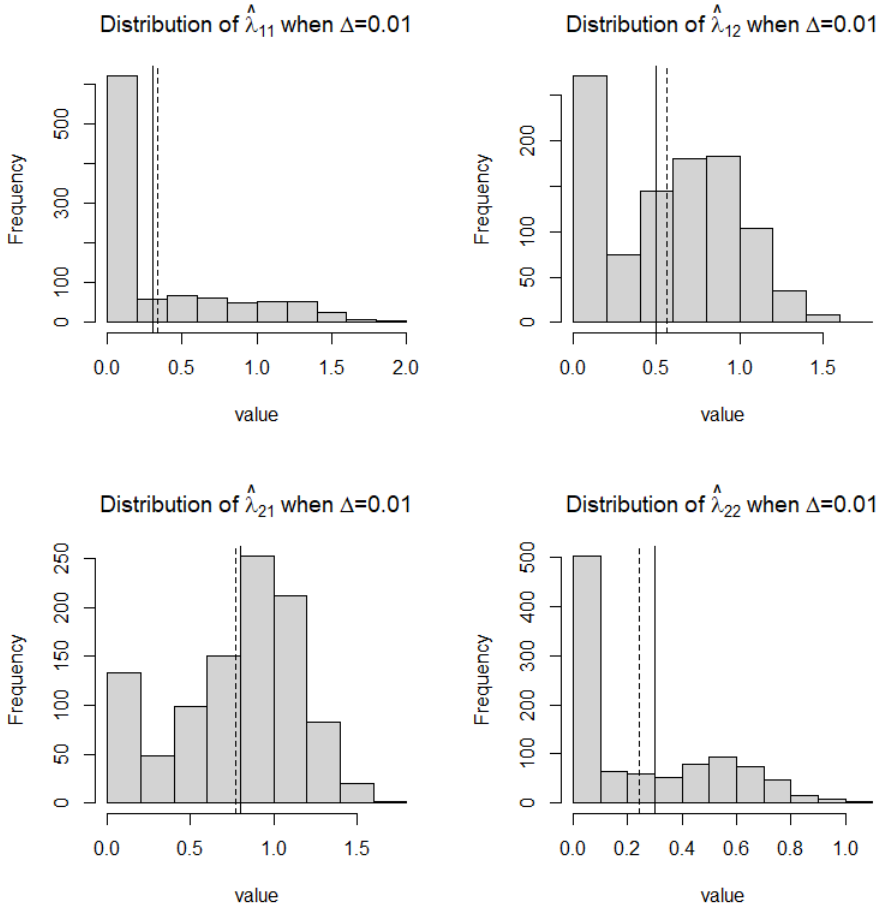


Figure 8.9: Empirical distribution for individual birth and death rates. The solid lines are the true values of the parameters listed in table 8.9 and the dash lines stand for empirical means.

8.6 Concluding remarks

In this paper, we propose an integer-valued autoregressive model INAR(1) to approximate the continuous birth-and-death process. In univariate case, we propose a birth-death operator $p *_1 \alpha \circ X$ which is the sum of zero-modified geometric random variable. The parametrization of p and α can be determined by matching the first and second moment of continuous process. Then we propose an bivariate INAR(1) model to approximate bivariate birth and death process where birth probabilities will also depend on the size of the other population. The parametrization of this model can be obtained in a similar way. The convergence from discrete process to continuous process is proved by apply weak convergence theorem of locally bounded semimartingales. Due to the simple Markov structure of INAR(1) model, maximum likelihood estimation would be feasible. It is however not the case for bivariate and

multivariate birth and death process. Basically, one can extend the result here to multivariate case, i.e. we can approximate multivariate birth and death process in Griffiths (1973) by multivariate INAR(1) model using the these operators $*_1, *_2, \circ$ only as well as adding an immigrant process. However, the difficulty of expressing the parameters of INAR(1) model in terms of the parameters of multivariate birth and death process would be increasing and as we need to find out the first moment of birth and process explicitly.

8.A Proof of proposition 8.1

Similar to univariate case defined in (8.6), the infinitesimal generator of the bivariate birth and death process $(M_{1,t}, M_{2,t})$ acting on a function $f(t, M_1, M_2)$ within its domain $\Omega(\mathcal{A})$ is given by

$$\begin{aligned}\mathcal{A}f(t, M_1, M_2) =& \frac{\partial f}{\partial t} + (\lambda_{11}M_1 + \lambda_{21}M_2)(f(t, M_1 + 1, M_2) - f(t, M_1, M_2)) \\ &+ \mu_1 M_1(f(t, M_1 - 1, M_2) - f(t, M_1, M_2)) \\ &+ (\lambda_{12}M_1 + \lambda_{22}M_2)(f(t, M_1, M_2 + 1) - f(t, M_1, M_2)) \\ &+ \mu_2 M_2(f(t, M_1, M_2 - 1) - f(t, M_1, M_2)),\end{aligned}$$

where $\Omega(\mathcal{A})$ is the domain for the generator $\mathcal{A}$ such that $f(t, M_1, M_2)$ is differentiable with respect to t for all t, M_1, M_2 and

$$\begin{aligned}&|f(t, M_1 + 1, M_2) - f(t, M_1, M_2)| < \infty, \quad |f(t, M_1 - 1, M_2) - f(t, M_1, M_2)| < \infty \\ &|f(t, M_1, M_2 + 1) - f(t, M_1, M_2)| < \infty, \quad |f(t, M_1, M_2 - 1) - f(t, M_1, M_2)| < \infty\end{aligned}$$

Apply infinitesimal generator $\mathcal{A}$ to functions $f(t, M_1, M_2) = M_{1,t}, M_{2,t}, M_{1,t}^2, M_{2,t}^2$ and $M_{1,t}M_{2,t}$ respectively, we have

$$\begin{aligned}\mathcal{A}M &= (\lambda_{11}M_1 + \lambda_{21}M_2)(M_1 + 1 - M_1) + \mu_1 M_1(M_1 - 1 - M_1) \\ \mathcal{A}N &= (\lambda_{12}M_1 + \lambda_{22}M_2)(M_2 + 1 - M_2) + \mu_2 M_2(M_2 - 1 - M_2)\end{aligned}$$

$$\begin{aligned}\mathcal{A}M_1^2 &= (\lambda_{11}M_1 + \lambda_{21}M_2)((M_1 + 1)^2 - M_1^2) + \mu_1 M_1((M_1 - 1)^2 - M_1^2) \\ \mathcal{A}M_2^2 &= (\lambda_{12}M_1 + \lambda_{22}M_2)((M_2 + 1)^2 - M_2^2) + \mu_2 N((M_2 - 1)^2 - M_2^2) \\ \mathcal{A}M_1M_2 &= (\lambda_{11}M_1 + \lambda_{21}M_2)((M_1 + 1)N - M_1M_2) + (\lambda_{12}M_1 + \lambda_{22}M_2)(M_1(M_2 + 1) - M_1M_2) \\ &\quad + \mu_1 M_1((M_1 - 1)M_2 - M_1M_2) + \mu_2(M_1(M_2 - 1) - M_1M_2)\end{aligned}$$

The first two result in the following system of ODE

$$\begin{aligned}\frac{d}{dt}\mathbb{E}[M_{1,t}] &= \lambda_{21}\mathbb{E}[M_{2,t}] - \kappa_1\mathbb{E}[M_{1,t}] \\ \frac{d}{dt}\mathbb{E}[M_{2,t}] &= \lambda_{12}\mathbb{E}[M_{1,t}] - \kappa_2\mathbb{E}[M_{2,t}]\end{aligned} \tag{8.60}$$

The other three equations become (8.17), which is hard to solve explicitly as we need to solve an inhomogeneous ordinary differential equation system. To solve the system (8.60), we can first assume a linear relationship between $\mathbb{E}[M_{1,t}]$ and $\mathbb{E}[M_{2,t}]$ such that $\mathbb{E}[M_{1,t}] = c\mathbb{E}[M_{2,t}] + g(t)$ for some constant c and a real value function g. Applying this substitution, the first ODE in (8.60) becomes

$$c\frac{d}{dt}\mathbb{E}[M_{2,t}] + g'(t) = \lambda_{21}\mathbb{E}[M_{2,t}] - \kappa_1(c\mathbb{E}[M_{2,t}] + g(t))$$

Make use of the second the ODE in , the first ODE can be rearranged into a ordinary equation.

$$\begin{aligned}
&c(\lambda_{12}\mathbb{E}[M_{1,t}] - \kappa_2\mathbb{E}[M_{2,t}]) + g'(t) = \lambda_{21}\mathbb{E}[M_{2,t}] - \kappa_1\mathbb{E}[M_{1,t}] \\
&(\lambda_{12}c + \kappa_1)\mathbb{E}[M_{1,t}] - (c\kappa_2 + \lambda_{21})\mathbb{E}[M_{2,t}] + g'(t) = 0 \\
&(\lambda_{12}c + \kappa_1)(c\mathbb{E}[M_{2,t}] + g(t)) - (c\kappa_2 + \lambda_{21})\mathbb{E}[M_{2,t}] + g'(t) = 0 \\
&(\lambda_{12}c^2 + (\kappa_1 - \kappa_2)c - \lambda_{21})\mathbb{E}[M_{2,t}] + g'(t) + (\lambda_{12} + \kappa_1)g(t) = 0
\end{aligned}$$

Then c is the solution of the quadratic equation

$$\begin{aligned}
&\lambda_{12}c^2 + (\kappa_1 - \kappa_2)c - \lambda_{21} = 0 \\
&c = \frac{\kappa_2 - \kappa_1 \pm \sqrt{(\kappa_1 - \kappa_2)^2 + 4\lambda_{21}\lambda_{12}}}{2\lambda_{12}}
\end{aligned}$$

Both roots would result in the same moments, so just take the positive root. The function g would be the solution of following ODE

$$\begin{aligned}
&g'(t) + (\lambda_{12} + \kappa_1)g(t) = 0 \\
&g(t) = g(0)e^{-(\lambda_{12}c - \kappa_2)t} \\
&g(0) = \mathbb{E}[M_{1,0}] - c\mathbb{E}[M_{2,0}] = (a - cb)
\end{aligned}$$

Then $\mathbb{E}[M_{2,t}]$ is determined by the following ODE

$$\frac{d}{dt}\mathbb{E}[M_{2,t}] = \lambda_{12}\mathbb{E}[M_{1,t}] - \kappa_2\mathbb{E}[M_{2,t}] = (\lambda_{12}c - \kappa_2)\mathbb{E}[M_{2,t}] + \lambda_{12}g(t)$$

and $\mathbb{E}[M_{1,t}] = c\mathbb{E}[M_{2,t}] + g(t)$ □

8.B Proof of proposition 8.2

For the first property, we can verify in the following way.
When $i = 1$

$$d_0 = p^0 \left(1 + (1-p) \frac{\frac{\alpha}{p} - \left(\frac{\alpha}{p}\right)}{1 - \frac{\alpha}{p}} \right) = 1$$

$$p_1 = p, \quad \alpha_1 = \alpha$$

Suppose equation (8.22) holds for $i = k$. Then for $i = k+1$, we have

$$\begin{aligned}
\phi^{(I)}(t,\theta) &= \mathbb{E}\left[\left(1 - \alpha_k + \frac{\alpha_k p_k \theta}{1-(1-p_k)\theta} \right)^{X_{t-k}} \right] \\
&= \mathbb{E}\left[\left(\frac{1 - \alpha_k - (1-\alpha_k - p_k)\theta}{1-(1-p_k)\theta} \right)^{X_{t-k}} \right] \\
&= \mathbb{E}\left[\left(\frac{1 - \alpha - (1-\alpha-p)\frac{1-\alpha_k-(1-\alpha_k-p_k)\theta}{1-(1-p_k)\theta}}{1-(1-p)\frac{1-\alpha_k-(1-\alpha_k-p_k)\theta}{1-(1-p_k)\theta}} \right)^{X_{t-k-1}} \right] \\
&= \mathbb{E}\left[\left(\frac{\frac{1-\alpha-(1-\alpha-p)(1-\alpha_k)}{1-(1-p)(1-\alpha_k)} - \left(\frac{(1-p_k+(1-p)(\alpha_k-p_k-1))}{1-(1-p)(1-\alpha_k)} - \frac{\alpha\alpha_k}{1-(1-p)(1-\alpha_k)} \right)\theta}{1 - \frac{(1-p_k+(1-p)(\alpha_k-p_k-1))}{1-(1-p)(1-\alpha_k)}\theta} \right)^{X_{t-k-1}} \right]
\end{aligned}$$

It is then clear that

$$\begin{aligned}
1-(1-p)(1-\alpha_k) &= \frac{d_{k-1} - (1-p)(d_{k-1} - \alpha^k)}{d_{k-1}} = \frac{d_k}{d_{k-1}} \\
\alpha_{k+1} &= \frac{\alpha\alpha_k}{1-(1-p)(1-\alpha_k)} = \frac{\alpha^{k+1}}{d_k} \\
p_{k+1} &= 1 - \frac{(1-p_k+(1-p)(\alpha_k-p_k-1))}{1-(1-p)(1-\alpha_k)} \\
&= 1 - \frac{1-(1-p)(1-\alpha_k) - pp_k}{1-(1-p)(1-\alpha_k)} = \frac{pp_k}{1-(1-p)(1-\alpha_k)} = \frac{p^{k+1}}{d_k}
\end{aligned}$$

So equation (8.22) holds for all $i = 1, 2, \ldots$. For the second property, the moments can be found by conditional expectation such that

$$\begin{aligned}
\mathbb{E}[p *_1 \alpha \circ X|X] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{\alpha\circ X} g_j^{(j)}|\alpha \circ X\right]|X\right] = \frac{1}{p}\mathbb{E}[(\alpha \circ X)|X] = \frac{1}{p}\mathbb{E}\left[\sum_{j=1}^{X} b_j|X\right] = \frac{\alpha}{p}X \\
Var(p *_1 \alpha \circ X|X) &= \mathbb{E}\left[Var\left(\sum_{j=1}^{\alpha\circ X} g_j^{(j)}|\alpha \circ X\right)|X\right] + Var\left(\mathbb{E}\left[\sum_{j=1}^{\alpha\circ X} g_j^{(j)}|\alpha \circ X\right]|X\right) \\
&= \frac{1-p}{p^2}\mathbb{E}[\alpha \circ X|X] + \frac{1}{p^2}Var(\alpha \circ X|X) \\
&= \left(\frac{\alpha(1-p)}{p^2} + \frac{\alpha(1-\alpha)}{p^2}\right)X \\
Var(p *_1 \alpha \circ X) &= \mathbb{E}[Var(p *_1 \alpha \circ X|X)] + Var(\mathbb{E}[p *_1 \alpha \circ X|X]) \\
Cov(X_t, X_{t-i}) &= Cov(p_i *_1 \alpha_i \circ X_{t-i}, X_{t-i}) \\
&= Cov\left(\mathbb{E}[p_i *_1 \alpha_i \circ X_{t-i}|X_{t-i}], \mathbb{E}[X_{t-i}|X_{t-i}]\right) \\
&\quad + \mathbb{E}[Cov(p_i *_1 \alpha_i \circ X_{t-i}, X_{t-i}|X_{t-i})] \\
&= \frac{\alpha_i}{p_i}Cov(X_{t-i}, X_{t-i}) + 0
\end{aligned}$$

□

8.C Proof of Proposition 8.5

The transition probability of giving out $m(m > 0)$ birth of from the process $Z_t^{(\Delta)}$ given $Z_{t-\Delta}^{(\Delta)}$ during an infinitesimal time interval Δ is given by

$$\begin{aligned}
&\Pr(Z_t^{(\Delta)} = k + m|Z_{t-\Delta}^{(\Delta)} = k) = \Pr(X_s = k + m|X_{s-1} = k) \\
&= \sum_{j=1}^{\min\{k,k+m\}} \binom{k}{j}\alpha^j(1-\alpha)^{k-j}\binom{k+m-1}{j-1}p^j(1-p)^{k+m-j} \\
&= \sum_{j=1}^{k} \binom{k}{j}\binom{k+m-1}{j-1}e^{(\lambda-\mu)j\Delta}\lambda^{k+m-j}\mu^{k-j}(e^{(\lambda-\mu)\Delta}-1)^{2k-2j+m}(\lambda-\mu)^{2j} \\
&\quad \times (\lambda e^{(\lambda-\mu)\Delta} - \mu)^{-(2k+m)} \\
&= \sum_{j=1}^{k-1} \binom{k}{j}\binom{k+m-1}{j-1}\lambda^{k-j}\mu^{k-j}(\lambda-\mu)^{2j}((\lambda-\mu)\Delta + o(\Delta))^{2k-2j+m} \\
&\quad \times (1 + (\lambda-\mu)\Delta + o(\Delta))^j \left(\frac{1}{\lambda-\mu} - \frac{\lambda}{\lambda-\mu}\Delta + o(\Delta)\right)^{2k+m},
\end{aligned}$$

where all the exponential function are expressed as their corresponding Taylor expansion at $\Delta = 0$. To make comparison with the continuous birth and death process, we are interested in the coefficients in front of Δ. First we need to check the lowest order of Δ in above probability. That is, we would like to minimize the sum

$$\min_{1 \leqslant j \leqslant k} 2k - 2j + m = m$$

So for the transition probability is rearranged in the following way

$$\begin{aligned}
&\Pr(Z_t^{(\Delta)} = k + m | Z_{t-\Delta}^{(\Delta)} = k) \\
=&\binom{k}{k}\binom{k+m-1}{k-1}\lambda^m(\lambda-\mu)^{2k}((\lambda-\mu)\Delta + o(\Delta))^m(1+(\lambda-\mu)\Delta+o(\Delta))^k \\
&\times \left(\frac{1}{\lambda-\mu} - \frac{\lambda}{\lambda-\mu}\Delta + o(\Delta)\right)^{2k+m} + o(\Delta^m) \\
=&\binom{k+m-1}{k-1}(\lambda\Delta)^m + o(\Delta^m),
\end{aligned}$$

Then it is clear that the there is no first order term in the case where $m \geqslant 2$ and the probability that giving out exactly one birth is

$$\Pr(Z_t^{(\Delta)} = k + 1 | Z_{t-\Delta}^{(\Delta)} = k) = \lambda k \Delta + o(\Delta)$$

On the other hand, we can derive the probability that $m (1 \leqslant m \leqslant k)$ individuals die within infinitesimal time Δ in a similar way

$$\begin{aligned}
&\Pr(Z_t^{(\Delta)} = k - m | Z_{t-\Delta}^{(\Delta)} = k) = \Pr(X_s = k - m | X_{s-1} = k) \\
=& \sum_{j=1}^{\min\{k-m,k\}} \binom{k}{j}\alpha^j(1-\alpha)^{k-j}\binom{k-m-1}{j-1}p^j(1-p)^{k-m-j} \\
=& \sum_{j=1}^{k-m} \binom{k}{j}\binom{k-m-1}{j-1}e^{(\lambda-\mu)j\Delta}\lambda^{k-m-j}\mu^{k-j}(e^{(\lambda-\mu)\Delta}-1)^{2k-2j-m} \\
&\times (\lambda-\mu)^{2j}(\lambda e^{(\lambda-\mu)\Delta} - \mu)^{-(2k-m)} \\
=& \sum_{j=1}^{k-m} \binom{k}{j}\binom{k-m-1}{j-1}\lambda^{k-m-j}\mu^{k-j}(\lambda-\mu)^{2j}((\lambda-\mu)\Delta+o(\Delta))^{2k-2j-m} \\
&\times (1+(\lambda-\mu)\Delta+o(\Delta))^j\left(\frac{1}{\lambda-\mu} - \frac{\lambda}{\lambda-\mu}\Delta + o(\Delta)\right)^{2k-m}
\end{aligned}$$

The minimum order of Δ is determined by

$$\min_{1 \leqslant j \leqslant k-m} 2k - 2j - m = m \tag{8.61}$$

Then the probability is reduced to

$$
\begin{aligned}
&\Pr(Z_t^{(\Delta)} = k-m | Z_{t-\Delta}^{(\Delta)} = k) \\
=&\binom{k}{k-m}\binom{k-m-1}{k-m-1}\mu^m(\lambda-\mu)^{2(k-m)}((\lambda-\mu)\Delta+o(\Delta))^m(1+(\lambda-\mu)\Delta+o(\Delta))^{k-m} \\
&\times\left(\frac{1}{\lambda-\mu}-\frac{\lambda}{\lambda-\mu}\Delta+o(\Delta)\right)^{2k-m}+o(\Delta^m) \\
=&\binom{k}{k-m}(\mu\Delta)^m+o(\Delta^m)
\end{aligned}
\tag{8.62}
$$

The transition probability that only one individual dies is

$$
\Pr(Z_t^{(\Delta)} = k-1 | Z_{t-\Delta}^{(\Delta)} = k) = \mu k\Delta + o(\Delta) \tag{8.63}
$$

the probability that more than one individuals die are $o(\Delta)$ As the birth rate and death (λ, μ) are time-homogeneous, so as the parameters α, p, the transition probabilities stay the same for all time $t \in [t_1, t_2]$. This means that the discrete birth and death INAR(1) model would result in the the same dynamic (8.1) of simple birth and death process when Δ is small enough.

Similar to the univariate case. It is necessary to find out the transition probabilities before proceeding to the weak convergence. The transition probability of giving out $m(m \geqslant 1)$ births of population $M_t^{(\Delta)}$ given $M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2$ during an

infinitesimal time Δ is given by

$$
\begin{aligned}
&\Pr(M_t^{(\Delta)} = k_1 + m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) = \Pr(X_s = k_1 + m | X_{s-1} = k_1, Y_{s-1} = k_2) \\
&= \sum_{j=0}^{k_1+m} \Pr(\beta_{11} *_1 \alpha_1 \circ X_{s-1} = j | X_{s-1} = k_1) \Pr(\beta_{21} *_2 Y_{s-1} = k_1 + m - j | Y_{s-1} = k_2) \\
&= \sum_{j=1}^{k_1+m} \left(\sum_{i=1}^{\min\{k_1,j\}} \binom{k_1}{i} \alpha_1^i (1-\alpha_1)^{k_1-i} \binom{j-1}{i-1} \beta_{11}^i (1-\beta_{11})^{j-i} \right) \binom{k_2 + k_1 + m - j - 1}{k_2 - 1} \\
&\quad \times \beta_{21}^{k_2} (1-\beta_{21})^{k_1+m-j} + (1-\alpha_1)^{k_1} \binom{k_2 + k_1 + m - 1}{k_2 - 1} \beta_{21}^{k_2} (1-\beta_{21})^{k_1+m} \\
&= \sum_{j=1}^{k_1+m} \sum_{i=1}^{\min\{k_1,j\}} \binom{k_1}{i} \binom{j-1}{i-1} \binom{k_2 + k_1 + m - j - 1}{k_2 - 1} \left(\frac{(\lambda_{11} - \mu_1)\omega_1(\Delta)}{\lambda_{11}\omega_1(\Delta) - \mu_1} \right)^i \\
&\quad \times \left(\frac{\mu_1(\omega_1(\Delta) - 1)}{\lambda_{11}\omega_1(\Delta) - \mu_1} \right)^{k_1-i} \left(\frac{\lambda_{11} - \mu_1}{\lambda_{11}\omega_1(\Delta) - \mu_1} \right)^i \left(\frac{\lambda_{11}(\omega_1(\Delta) - 1)}{\lambda_{11}\omega_1(\Delta) - \mu_1} \right)^{j-i} \\
&\quad \times \left(\frac{1}{1 + C_{\beta_1}(e^{u_1\Delta} - e^{u_2\Delta})} \right)^{k_2} \left(\frac{C_{\beta_1}(e^{u_1\Delta} - e^{u_2\Delta})}{1 + C_{\beta_1}(e^{u_1\Delta} - e^{u_2\Delta})} \right)^{k_1+m-j} + o(\Delta^{2k_1+m-1}) \\
&= \sum_{j=1}^{k_1+m} \sum_{i=1}^{\min\{k_1,j\}} \binom{k_1}{i} \binom{j-1}{i-1} \binom{k_2 + k_1 + m - j - 1}{k_2 - 1} (\lambda_{11} - \mu_1)^i (1 + (\lambda_{11} - \mu_1)\Delta + o(\Delta))^i \\
&\quad \times \mu_1^{k_1-i} ((\lambda_{11} - \mu_1)\Delta + o(\Delta))^{k_1-i} (\lambda_{11} - \mu_1)^i \lambda_{11}^{j-i} ((\lambda_{11} - \mu_1)\Delta + o(\Delta))^{j-i} \\
&\quad \times \left(\frac{1}{\lambda_{11} - \mu_1} - \frac{\lambda_{11}}{\lambda_{11} - \mu_1}\Delta + o(\Delta) \right)^{k_1+j} \\
&\quad \times (1 - C_{\beta_1}(u_1 - u_2)\Delta + o(\Delta))^{k_2} (C_{\beta_1}(u_1 - u_2)\Delta + o(\Delta))^{k_1+m-j} + o(\Delta^{2k_1+m-1}) \\
&= \sum_{j=1}^{k_1+m} \sum_{i=1}^{\min\{k_1,j\}} \binom{k_1}{i} \binom{j-1}{i-1} \binom{k_2 + k_1 + m - j - 1}{k_2 - 1} (\lambda_{11} - \mu_1)^{2i} \lambda_{11}^{j-i} \mu_1^{k_1-i} \\
&\quad \times ((\lambda_{11} - \mu_1)\Delta + o(\Delta))^{k_1+j-2i} (1 + (\lambda_{11} - \mu_1)\Delta + o(\Delta))^i \\
&\quad \times \left(\frac{1}{\lambda_{11} - \mu_1} - \frac{\lambda_{11}}{\lambda_{11} - \mu_1}\Delta + o(\Delta) \right)^{k_1+j} \\
&\quad \times (1 - C_{\beta_1}(u_1 - u_2)\Delta + o(\Delta))^{k_2} (C_{\beta_1}(u_1 - u_2)\Delta + o(\Delta))^{k_1+m-j} + o(\Delta^{2k_1+m-1}),
\end{aligned}
$$

where all the exponential function are expressed in their corresponding Taylor expansion at $\Delta = 0$. The lowest order of Δ is determined by the power of $((\lambda_{11} - \mu_1) + o(\Delta))$ and $(C_{\beta_1}(u_1 - u_2)\Delta + o(\Delta))$,

$$
\min_{1 \leqslant i \leqslant \min\{k_1,j\}} k_1 + j - 2i + k_1 + m - j = \min_{1 \leqslant i \leqslant \min\{k_1,j\}} 2k_1 - 2i + m = m,
$$

where $j = \in \{1, \ldots, k_1 + m\}$. This leads to $j = k_1, i = k_1 - 1$ and $j = k_1 - 1, i = k_1 - 1$,

respectively. Then the transition probability reduces to

$$
\begin{aligned}
&\Pr(M_t^{(\Delta)} = k_1 + m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \\
&= \sum_{j=k_1}^{k_1+m} \binom{k_1}{k_1} \binom{j-1}{k_1-1} \binom{k_2+k_1+m-j-1}{k_2-1} (\lambda_{11}-\mu_1)^{2k_1} \lambda_{11}^{j-k_1} \mu_1^{k_1-k_1} \\
&\quad \times ((\lambda_{11}-\mu_1)\Delta + o(\Delta))^{k_1+j-2k_1} (1 + (\lambda_{11}-\mu_1)\Delta + o(\Delta))^{k_1} \\
&\quad \times \left(\frac{1}{\lambda_{11}-\mu_1} - \frac{\lambda_{11}}{\lambda_{11}-\mu_1}\Delta + o(\Delta) \right)^{k_1+j} \\
&\quad \times (1 - C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_2} (C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_1+m-j} + o(\Delta^{2k_1+m-1}), \\
&= \sum_{j=k_1}^{k_1+m} \binom{j-1}{k_1-1} \binom{k_2+k_1+m-j-1}{k_2-1} (\lambda_{11}\Delta)^{j-k_1} (C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_1+m-j} \\
&\quad + o(\Delta^{2k_1+m-1}), \\
&= \sum_{j=k_1}^{k_1+m} \binom{j-1}{k_1-1} \binom{k_2+k_1+m-j-1}{k_2-1} (\lambda_{11}\Delta)^{j-k_1} (\lambda_{21}\Delta)^{k_1+m-j} + o(\Delta^m),
\end{aligned}
$$

Then the probability when $m = 1$ is given by

$$
\begin{aligned}
&\Pr(M_t^{(\Delta)} = k_1 + 1 | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \\
&= \sum_{j=k_1}^{k_1+1} \binom{j-1}{k_1-1} \binom{k_2+k_1-j}{k_2-1} (\lambda_{11}\Delta)^{j-k_1} (\lambda_{21}\Delta)^{k_1+1-j} + o(\Delta), \qquad (8.64) \\
&= k_2\lambda_{21}\Delta + k_1\lambda_{11}\Delta + o(\Delta)
\end{aligned}
$$

By symmetry, the birth transition probability for the other population is given by

$$
\begin{aligned}
&\Pr(N_t^{(\Delta)} = k_2 + m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) = P(Y_s = k_2 | X_{s-1} = k_1, Y_{s-1} = k_2) \\
&= \sum_{j=k_2}^{k_2+m} \binom{j-1}{k_2-1} \binom{k_1+k_2+m-j-1}{k_1-1} (\lambda_{22}\Delta)^{j-k_2} (\lambda_{12}\Delta)^{k_1+m-j} + o(\Delta^m).
\end{aligned}
$$

On the other hand, the probability that m individual die in population $M_t^{(\Delta)}$ given

that $M_{t-\Delta}^{(\Delta)} = k_1$, $N_{t-\Delta}^{(\Delta)} = k_2$ during an infinitesimal time interval Δ is

$$
\begin{aligned}
&\Pr(M_t^{(\Delta)} = k_1 - m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) = \Pr(X_s = k_1 - m | X_{s-1} = k_1, Y_{s-1} = k_2) \\
=& \sum_{j=1}^{k_1-m} \left(\sum_{i=1}^{\min\{k_1,j\}} \binom{k_1}{i} \alpha_1^i (1-\alpha_1)^{k_1-i} \binom{j-1}{i-1} \beta_{11}^i (1-\beta_{11})^{j-i} \right) \\
&\times \binom{k_2+k_1-m-j-1}{k_2-1} \beta_{21}^{k_2} (1-\beta_{21})^{k_1-j} + (1-\alpha_1)^{k_1} \binom{k_2+k_1-m-1}{k_2-1} \beta_{21}^{k_2} (1-\beta_{21})^{k_1} \\
=& \sum_{j=1}^{k_1-m} \sum_{i=1}^{j} \binom{k_1}{i} \binom{j-1}{i-1} \binom{k_2+k_1-m-j-1}{k_2-1} (\lambda_{11}-\mu_1)^{2i} \lambda_{11}^{j-i} \mu_1^{k_1-i} \\
&\times ((\lambda_{11}-\mu_1)\Delta + o(\Delta))^{k_1+j-2i} \\
&\times (1 + (\lambda_{11}-\mu_1)\Delta + o(\Delta))^i \left(\frac{1}{\lambda_{11}-\mu_1} - \frac{\lambda_{11}}{\lambda_{11}-\mu_1}\Delta + o(\Delta) \right)^{k_1+j} \\
&\times (1 - C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_2} (C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_1-m-j} + o(\Delta^{2k_1+k_2-1}) \\
=& \binom{k_1+1}{k_1} \binom{k_1-1}{k_1-1} \binom{k_2-1}{k_2-1} (\lambda_{11}-\mu_1)^{2k_1} \mu_1 ((\lambda_{11}-\mu_1)\Delta + o(\Delta)) \\
&\times (1 + (\lambda_{11}-\mu_1)\Delta + o(\Delta))^{k_1} \\
&\times \left(\frac{1}{\lambda_{11}-\mu_1} - \frac{\lambda_{11}}{\lambda_{11}-\mu_1}\Delta + o(\Delta) \right)^{2k_1+1} (1 - C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_2} + o(\Delta) \\
=& \mu_1(k_1+1)\Delta + o(\Delta)
\end{aligned}
$$

The lowest order of Δ is determined by the power of $((\lambda_{11}-\mu_1)+o(\Delta))$ and $(C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))$,which is

$$\min_{1 \leqslant i \leqslant j} k_1 + j - 2i + k_1 - m - j = \min_{1 \leqslant i \leqslant j} 2k_1 - 2i - m = m,$$

where $j \in \{1, \ldots, k_1 - m\}$. So the above probability reduces to

$$
\begin{aligned}
&\Pr(M_t^{(\Delta)} = k_1 - m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \\
=& \binom{k_1}{k_1-m} \binom{k_1-m-1}{k_1-m-1} \binom{k_2-1}{k_2-1} (\lambda_{11}-\mu_1)^{2(k_1-m)} \mu_1^m \\
&\times ((\lambda_{11}-\mu_1)\Delta + o(\Delta))^m \\
&\times (1 + (\lambda_{11}-\mu_1)\Delta + o(\Delta))^{k_1-m} \left(\frac{1}{\lambda_{11}-\mu_1} - \frac{\lambda_{11}}{\lambda_{11}-\mu_1}\Delta + o(\Delta) \right)^{2k_1-m} \\
&\times (1 - C_{\beta_1}(u_1-u_2)\Delta + o(\Delta))^{k_2} + o(\Delta^m) \\
=& \binom{k_1}{k_1-m} (\mu_1\Delta)^m + o(\Delta^m)
\end{aligned}
$$

It is not surprise that the transition probability only depends on its own size of population from the bivariate INAR construction. Then the death transition probability for the other population is

$$\Pr(N_t^{(\Delta)} = k_2 - m | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) = \binom{k_2}{k_2 - m} (\mu_2 \Delta)^m + o(\Delta^m) \quad (8.65)$$

The it is clear that the probabilities for both population that there is only one death would have the same form as in the univariate case. By conditional independence of bivariate INAR model, the joint transition probability would be the product of any of two transition probabilities shown above. For example, for any two integers $m_1, m_2 \in \mathbb{Z}$,

$$\begin{aligned} &\Pr(M_t^{(\Delta)} = k_1 + m_1, N_t^{(\Delta)} = k_2 + m_2 | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \\ =&\Pr(M_t^{(\Delta)} = k_1 + m_1 | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \Pr(N_t^{(\Delta)} = k_2 + m_2 | M_{t-\Delta}^{(\Delta)} = k_1, N_{t-\Delta}^{(\Delta)} = k_2) \end{aligned} \quad (8.66)$$

Then it is straightforward to show that, to have a first order Δ term, the only possible combinations of (m_1, m_2) are $\{(1,0),(0,1),(-1,0),(0,-1)\}$, which under the proposed parametrization, the joint process only allow one jump during infinitesimal time which coincide with the bivariate continuous birth and death process.

As the birth rates $\lambda_{i,j}$ and death rate are μ_i time homogeneous, so as the parameters $\alpha_i, \beta_{i,j}, i, j \in \{1,2\}$, the transition probabilities stay the same for all time $t \in [0,1]$. This means that this bivariate birth and death INAR(1) model would result in the the same dynamic (8.11) when Δ is small enough. □

8.D Proof of Proposition 8.6

According to Theorem 3.11, Chapter 2 in Jacod and Shiryaev (2013). We need to make sure the following two sum is finite for any truncation function h before constructing their discrete Lévy measures.

$$\begin{aligned} &\sum_{k=1}^{\sigma_t} |\mathbb{E}[h(U_k)|\mathcal{F}_{k-1}]| < \infty \\ &\sum_{k=1}^{\sigma_t} \mathbb{E}[|U_k^2 \wedge 1| \ |\mathcal{F}_{k-1}] < \infty, \end{aligned} \quad (8.67)$$

where U_k are increments of any underlying processes. This can be shown straightforwardly as the there are only finite number of terms for summation $\sigma_t \leqslant n$ and truncation functions are bounded. Then by The Theorem 3.11, the Lévy triplets for $Z_t^{(n)}$ is

$$Ch(Z_t^{(n)}) = \begin{cases} B_t = \sum_{k=1}^{\sigma_t} \mathbb{E}[h(U_k)|\mathcal{F}_{k-1}] \\ C_t = 0 \\ \nu(Z_t^{(n)}; [0,t] \times g) = \sum_{k=1}^{\sigma_t} \mathbb{E}[g(U_k)1_{\{U_k \neq 0\}})|\mathcal{F}_{k-1}] \end{cases} \tag{8.68}$$

We can choose the truncation function $h(x) = |x|1_{\{|x|<1\}}$ such that B_t is always 0 as there is no jump with size smaller than 1 in $Z_t^{(n)}$. Finally for the discrete stochastic integral:

$$\begin{aligned} &\mathbb{E}[g(U_k)1_{\{U_k \neq 0\}}|\mathcal{F}_{k-1}] \\ =&\mathbb{E}[g(1)1_{\{U_k=1\}}|\mathcal{F}_{k-1}] + \mathbb{E}[g(-1)1_{\{U_k=-1\}}|\mathcal{F}_{k-1}] \\ &+ \sum_{\eta=2}^{\infty} \mathbb{E}[g(\eta)1_{\{U_k=\eta\}} + g(-\eta)1_{\{U_k=-\eta\}}|\mathcal{F}_{k-1}] \\ =&g(1)\lambda X_{k-1}\Delta + g(-1)\mu X_{k-1}\Delta + o(\Delta) + \frac{o(\Delta^2)}{1-\Delta} \\ =&g(1)\lambda X_{k-1}\Delta + g(-1)\mu X_{k-1}\Delta + o(\Delta) \end{aligned} \tag{8.69}$$

Then it is clear that

$$\nu(Z_t^{(n)}; [0,t] \times g) = \sum_{k=1}^{\sigma_t} (g(1)\lambda + g(-1)\mu)X_{k-1}\Delta + O(\Delta)$$

For the bivariate case, the proof is similar, the conditional expectation

$$
\begin{aligned}
&\mathbb{E}[g(\mathbf{V}_k)1_{\{\mathbf{V}_k\neq(0,0)^T\}}|\mathcal{F}_{k-1}] \\
=&\mathbb{E}[g(1,0)1_{\{\mathbf{V}_k=(1,0)^T\}}|\mathcal{F}_{k-1}] + \mathbb{E}[g(0,1)1_{\{\mathbf{V}_k=(0,1)^T\}}|\mathcal{F}_{k-1}] \\
&+ \mathbb{E}[g(-1,0)1_{\{\mathbf{V}_k=(-1,0)^T\}}|\mathcal{F}_{k-1}] + \mathbb{E}[g(0,-1)1_{\{\mathbf{V}_k=(0,-1)^T\}}|\mathcal{F}_{k-1}] \\
&+ \sum_{|i|+|j|>1} \mathbb{E}[g(i,j)1_{\{\mathbf{V}_k=(i,j)^T\}}|\mathcal{F}_{k-1}] + \mathbb{E}[g(i,-j)1_{\{\mathbf{V}_k=(i,-j)^T\}}|\mathcal{F}_{k-1}] \\
&+ \sum_{|i|+|j|>1} \mathbb{E}[g(-i,j)1_{\{\mathbf{V}_k=(-i,j)^T\}}|\mathcal{F}_{k-1}] + \mathbb{E}[g(-i,-j)1_{\{\mathbf{V}_k=(-i,-j)^T\}}|\mathcal{F}_{k-1}] \\
=&\left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(0,-1)\tilde{\boldsymbol{\mu}}_2\right)\mathbf{Y}_{k-1}\Delta + o(\Delta) + \frac{o(\Delta^2)}{(1-\Delta)^2} \\
=&\left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(0,-1)\tilde{\boldsymbol{\mu}}_2\right)\mathbf{Y}_{k-1}\Delta + o(\Delta)
\end{aligned}
\tag{8.70}
$$

Finally the discrete stochastic integral is given by

$$
\sum_{k=1}^{\sigma_t}\left(g(1,0)\tilde{\boldsymbol{\lambda}}_1 + g(-1,0)\tilde{\boldsymbol{\mu}}_1 + g(0,1)\tilde{\boldsymbol{\lambda}}_2 + g(0,-1)\tilde{\boldsymbol{\mu}}_2\right)\mathbf{Y}_{k-1}\Delta + O(\Delta) \tag{8.71}
$$

□

CHAPTER 9

Concluding Remarks and Future Research

This thesis primarily investigates the realm of integer-valued time series and their linkage with classical point processes, specifically Poisson point processes, and birth-and-death processes. Focusing on the autoregressive structure within integer-valued time series, the first two papers (A, B) explore and develop different structure of univariate and multivariate INAR(1) model and introduce mixed Poisson random variables in innovations. This novel integration of mixed Poisson random variables imparts enhanced flexibility to the model, allowing it to accommodate varying levels of dispersion and correlation within and across count data sequences. Moreover, paper C develops an expectation and maximization algorithm to facilitate maximum likelihood estimation methods for previous proposed multivariate INAR(1) model. This development addresses the complexities in the likelihood function arising from the binomial thinning (autoregressive component) and diverse mixing densities of the mixed Poisson variables.

Shifting focus to the classical point processes, Papers D and E endeavor to discover discrete analogs to continuous point processes. Paper D introduces so called INARMA with infinity orders, aiming to approximate Poisson point processes The adoption of infinite orders in this model serves to approximate the decay function observed in the intensity of point processes. Similarly, Paper E proposes an integer-valued model tailored to approximate birth-and-death process, where the 'birth' and 'death' events are explicitly modelled through the interplay of Bernoulli and Geometric random variables. In summary, the rationale behind these approximation are three-folds: (i) Integer-valued models offer explicit formulations of point processes, enhancing interpretability. (ii) Many datasets lack precise time records for each

event, typically aggregating data at regular intervals instead. (iii) Integer-valued models present an alternative methodology for estimating point processes.

9.1 Application of Point Processes

The last two papers focus on theoretical development. However, exploring the practical applications of these stochastic processes is equally significant. Consider the Hawkes process, which can be conceptualized as an autoregressive Poisson point process, enjoy popularity for a long time due to many available inference methods. The Hawkes process finds diverse applications in finance, epidemiology, sociology and seismology. In contrast, the dynamic contagion process, which can be regarded as AMRA point process, received less attention, primarily due to the lack of robust inference methods. Nevertheless, it's important to highlight the distinct advantage of the dynamic contagion process over the Hawkes process, whose immigrant intensity ν in equation 7.5 is fixed or deterministic over time. Dynamic contagion process is useful when the underlying intensity is driven by another independent random event stream. For example, it is reasonable to adopt dynamic contagion process in epidemic (COVID-19) modelling Chen et al. (2021) as from general point of view, the transmission of a disease within a certain area is likely caused by a external event, e.g. virus carriers from overseas. In insurance modelling, Dassios and Zhao (2017b) considered a risk process with the arrival of claims modelled by a dynamic contagion process. Jang and Oh (2020) introduced a bivariate compound dynamic contagion process for the modelling of aggregate losses from cyber events.

Birth and death processes is relatively simple process. It is a continuous-time Markov chain that models a non-negative integer number of particles in a system. For the univariate one, it has been used extensively in many applications including evolution biology, ecology, population genetics, epidemiology, and queueing theory. Many applied models require the consideration of two or more interacting populations simultaneously to model behavior such as competition, predation, or infection and so on. Griffiths (1972) proposed bivariate birth-death process to model infectious disease Malaria. Xu et al. (2015) consider multi-typed birh-death-shift process to model evolution of mobile genetic elements. More recently, DeWitt et al. (2023) consider multi-typed birth and death process to model the phylodynamics and Azizi and Salari (2023) applies bivariate birth/death process for condition-based maintenance scheduling for a continuously monitored manufacturing system.

9.2 Perspectives on Future Research

The ideas brought forward in this thesis open the door to further research projects on integer-valued time series.

1. In the multivariate INAR(1) model, one can set some off-diagonal element of the matrix to be non-zero in equation 5.1 to introduce further cross correlation among the sequences. The challenge here will be the the evaluation of likelihood function as the non-zero off-diagonal elements will introduce further discrete convolution in probability transition functions. On the other hand, one can explore EM algorithm by introducing further latent variables from binomial parts.

2. On insurance claim regression modelling, based on the multivariate INAR(1) model proposed in Paper B, one can introduce further heterogeneity by applying a regression structure on ϕ such that $g(\phi) = Z'\beta$ where $g(.)$ is a link function and Z' can be a subset of design matrix of $Z = (z_1, z_2, \ldots, z_n)^T$, see the description following the equation 5.2. The choice of link function will depend on the range of parameter ϕ.

3. The EM algorithm proposed in paper C still require evaluation of transition probability in equation 6.22 which always involves discrete convolution and integrals. One can regard the mixing density function $f_{\boldsymbol{\phi}}(\theta)$ as a prior for θ and when $f_{\boldsymbol{\phi}}(\theta)$ is not a conjugate prior for gamma density $\eta(\boldsymbol{\theta}|\boldsymbol{\lambda}_t, \mathbf{k}_t)$, the integrals inside the probability transition function have to be evaluated through numerical methods. In multivariate setting, the probability transition function need a stable and fast computation method for discrete convolutions and multivariate integrals. One can address this challenge or find a way to avoid the computation of transition probabilities, e.g. variational inference method.

4. Unlike the autoregressive INAR(p), the likelihood functions for INMA and INARMA are difficult to construct as they involve many unobserved variables, it is the same problem for Cox process and dynamic contagion processes. It would be interesting to explore whether there are parametric or non-parametric ways to estimate INMA(p) models and INARMA(p, q) models so that these classical point process can be applied straightforwardly to the real data. Non-parametric methods (conditional least square) for INAR(p) model was purposed in Kirchner (2017).

5. Similarly, it will be interesting to seek another inference method other than maximum likelihood estimation for the birth-and-death INAR models proposed in Paper E as the likelihood function is already cumbersome in bivariate case. While I believe the INAR approximation can be well extent to multivariate linear birth and death models, fast inference methods are required to ensure applicability for such models.

Bibliography

Abdallah, A., Boucher, J.-P., and Cossette, H. (2016). Sarmanov family of multivariate distributions for bivariate dynamic claim counts model. *Insurance: Mathematics and Economics*, 68:120–133.

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606.

Al-Osh, M. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Al-Osh, M. and Alzaid, A. A. (1988a). First-order integer-valued autoregressive (INAR(1)) process: distributional and regression properties. *Statistica Neerlandica*, 42(1):53–61.

Al-Osh, M. and Alzaid, A. A. (1988b). Integer-valued moving average (INMA) process. *Statistical Papers*, 29(1):281–300.

Amalia, J., Purhadi, P., and Otok, B. W. (2017). Parameter estimation and statistical test of geographically weighted bivariate poisson inverse gaussian regression models. In *AIP Conference Proceedings*, volume 1905. AIP Publishing.

Atkinson, A. and Yeh, L. (1982). Inference for sichel's compound poisson distribution. *Journal of the American Statistical Association*, 77(377):153–158.

Azizi, F. and Salari, N. (2023). A novel condition-based maintenance framework for parallel manufacturing systems based on bivariate birth/birth–death processes. *Reliability Engineering & System Safety*, 229:108798.

Bacry, E., Dayri, K., and Muzy, J.-F. (2012). Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013a). Modelling microstructure noise with mutually exciting point processes. *Quantitative finance*, 13(1):65–77.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013b). Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Bailey, N. T. (1991). The elements of stochastic processes with applications to the natural sciences. 25.

Barndorff-Nielsen, O., Blaesild, P., and Seshadri, V. (1992). Multivariate distributions with generalized inverse gaussian marginals, and associated poisson mixtures. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 109–120.

Bartlett, M. S. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 25(2):264–281.

Bening, V. E. and Korolev, V. Y. (2012). *Generalized Poisson models and their applications in insurance and finance*. Walter de Gruyter.

Bermúdez, L., Guillén, M., and Karlis, D. (2018). Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance: Mathematics and Economics*, 83:161–169.

Bermúdez, L. and Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, 48(2):226–236.

Bermúdez, L. and Karlis, D. (2012). A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis*, 56(12):3988–3999.

Bermúdez, L. and Karlis, D. (2017). A posteriori ratemaking using bivariate poisson models. *Scandinavian Actuarial Journal*, 2017(2):148–158.

Bermúdez, L. and Karlis, D. (2021). Multivariate INAR (1) regression models based on the Sarmanov distribution. *Mathematics*, 9(5):505.

Bolancé, C., Guillen, M., and Pitarque, A. (2020). A Sarmanov distribution with beta marginals: An application to motor insurance pricing. *Mathematics*, 8(11).

Bolancé, C. and Vernic, R. (2019). Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution. *Insurance: Mathematics and Economics*, 85:89–103.

Boucher, J.-P., Denuit, M., and Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):135–162.

Bourguignon, M., Rodrigues, J., and Santos-Neto, M. (2019). Extended poisson INAR (1) processes with equidispersion, underdispersion and overdispersion. *Journal of Applied Statistics*, 46(1):101–118.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

Brännäs, K. and Hall, A. (2001). Estimation in integer-valued moving average models. *Applied Stochastic Models in Business and Industry*, 17(3):277–291.

Brännäs, K., Hellström, J., and Nordström, J. (2002). A new approach to modelling and forecasting monthly guest nights in hotels. *International Journal of Forecasting*, 18(1):19–30.

Bu, R., McCabe, B., and Hadri, K. (2008). Maximum likelihood estimation of higher-order integer-valued autoregressive processes. *Journal of Time Series Analysis*, 29(6):973–994.

Cameron, A. C., Li, T., Trivedi, P. K., and Zimmer, D. M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal*, 7(2):566–584.

Chen, R. and Hyrien, O. (2011). Quasi-and pseudo-maximum likelihood estimators for discretely observed continuous-time markov branching processes. *Journal of statistical planning and inference*, 141(7):2209–2227.

Chen, Z., Dassios, A., Kuan, V., Lim, J. W., Qu, Y., Surya, B., and Zhao, H. (2021). A two-phase dynamic contagion model for covid-19. *Results in Physics*, 26:104264.

Cheon, S., Song, S. H., and Jung, B. C. (2009). Tests for independence in a bivariate negative binomial model. *Journal of the Korean Statistical Society*, 38(2):185–190.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157.

Crawford, F. W., Minin, V. N., and Suchard, M. A. (2014). Estimation for general birth-death processes. *Journal of the American Statistical Association*, 109(506):730–747.

Daley, D. J. and Vere-Jones, D. (2003). An introduction to the theory of point processes, volume 1: Elementary theory and methods. *Verlag New York Berlin Heidelberg: Springer*.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Dassios, A. and Jang, J. (2008). The distribution of the interval between events of a cox process with shot noise intensity. *Journal of Applied Mathematics and Stochastic Analysis*, 2008:1–14.

Dassios, A. and Jang, J.-W. (2003). Pricing of catastrophe reinsurance and derivatives using the cox process with shot noise intensity. *Finance and Stochastics*, 7(1):73–95.

Dassios, A. and Jang, J. W. (2005). Kalman-bucy filtering for linear systems driven by the cox process with shot noise intensity and its application to the pricing of reinsurance contracts. *Journal of applied probability*, 42(1):93–107.

Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Advances in applied probability*, 43(3):814–846.

Dassios, A. and Zhao, H. (2017a). Efficient simulation of clustering jumps with CIR intensity. *Operations Research*, 65(6):1494–1515.

Dassios, A. and Zhao, H. (2017b). A generalized contagion process with an application to credit risk. *International Journal of Theoretical and Applied Finance*, 20(01):1750003.

Davis, M. H. (1984). Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):353–376.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N. (2016a). *Handbook of discrete-valued time series*. CRC Press.

Davis, R. A., Zang, P., and Zheng, T. (2016b). Sparse vector autoregressive modelling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

Davison, A. C., Hautphenne, S., and Kraus, A. (2021). Parameter estimation for discretely observed linear birth-and-death processes. *Biometrics*, 77(1):186–196.

Denuit, M., Guillen, M., and Trufin, J. (2019). Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, 13(2):378–399.

DeWitt, W. S., Evans, S. N., Hiesmayr, E., and Hummel, S. (2023). Mean-field interacting multi-type birth-death processes with a view to applications in phylodynamics. *arXiv preprint arXiv:2307.06010*.

Embrechts, P., Liniger, T., and Lin, L. (2011). Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378.

Feller, W. (1939). Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5:441–470.

Freeland, R. K. and McCabe, B. P. (2004). Analysis of low count time series data by poisson autoregression. *Journal of time series analysis*, 25(5):701–722.

Frees, E. W., Lee, G., and Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1):4.

Frey, S. and Sandås, P. (2009). The impact of iceberg orders in limit order books. In *AFA 2009 San Francisco Meetings Paper*.

Fung, T. C., Badescu, A. L., and Lin, X. S. (2019). A class of mixture of experts models for general insurance: Application to correlated claim frequencies. *ASTIN Bulletin: The Journal of the IAA*, 49(3):647–688.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

Ghitany, M., Karlis, D., Al-Mutairi, D., and Al-Awadhi, F. (2012). An em algorithm for multivariate mixed poisson regression models and its application. *Applied Mathematical Sciences*, 6(137):6843–6856.

Gómez-Déniz, E. and Calderín-Ojeda, E. (2021). A priori ratemaking selection using multivariate regression models allowing different coverages in auto insurance. *Risks*, 9(7):137.

Griffiths, D. (1972). A bivariate birth-death process which approximates to the spread of a disease involving a vector. *Journal of Applied Probability*, 9(1):65–75.

Griffiths, D. (1973). Multivariate birth-and-death processes as approximations to epidemic processes. *Journal of Applied Probability*, 10(1):15–26.

Gurmu, S. and Elder, J. (2000). Generalized bivariate count data regression models. *Economics Letters*, 68(1):31–36.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503.

Homburg, A., Weiß, C. H., Alwan, L. C., Frahm, G., and Göb, R. (2019). Evaluating approximate point forecasting of count processes. *Econometrics*, 7(3):30.

Jacobs, P. A. and Lewis, P. A. (1978a). Discrete time series generated by mixtures. i: Correlational and runs properties. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):94–105.

Jacobs, P. A. and Lewis, P. A. (1978b). Discrete time series generated by mixtures ii: asymptotic properties. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):222–228.

Jacobs, P. A. and Lewis, P. A. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36.

Jacod, J. and Shiryaev, A. (2013). Limit theorems for stochastic processes. 288.

Jang, J. and Oh, R. (2020). A bivariate compound dynamic contagion process for cyber insurance. *arXiv preprint arXiv:2007.04758*.

Jarrow, R. A. and Yu, F. (2001). Counterparty risk and the pricing of defaultable securities. *the Journal of Finance*, 56(5):1765–1799.

Jeong, H. and Dey, D. K. (2021). Multi-peril frequency credibility premium via shared random effects. *Available at SSRN 3825435*.

Jeong, H., Tzougas, G., and Fung, T. C. (2023). Multivariate claim count regression model with varying dispersion and dependence parameters. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(1):61–83.

Jin-Guan, D. and Yuan, L. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of time series analysis*, 12(2):129–142.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, volume 1*, volume 289. John wiley & sons.

Jung, R. C. and Tremayne, A. (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, 95(1):59–91.

Karlis, D. (2005). Em algorithm for mixed poisson and other discrete distributions. *ASTIN Bulletin: The Journal of the IAA*, 35(1):3–24.

Karlis, D. and Pedeli, X. (2013). Flexible bivariate INAR(1) processes using copulas. *Communications in Statistics-Theory and Methods*, 42(4):723–740.

Keiding, N. (1975). Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2):363–372.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282.

Kim, H. (2011). *Spatio-temporal point process models for the spread of avian influenza virus (H5N1)*. PhD thesis, UC Berkeley.

Kirchner, M. (2016). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525.

Kirchner, M. (2017). An estimation procedure for the hawkes process. *Quantitative Finance*, 17(4):571–595.

Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1–25.

Latour, A. (1997). The multivariate ginar (p) process. *Advances in Applied Probability*, pages 228–248.

Lee, A. (1999). Applications: Modelling rugby league data viabivariate negative binomial regression. *Australian & New Zealand Journal of Statistics*, 41(2):141–152.

Lee, G. Y. and Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, 87:115–129.

Mardalena, S., Purhadi, P., Purnomo, J. D. T., and Prastyo, D. D. (2020). Parameter estimation and hypothesis testing of multivariate poisson inverse gaussian regression. *Symmetry*, 12(10):1738.

Mardalena, S., Purnomo, J., Prastyo, D., et al. (2021). Bivariate poisson inverse gaussian regression model with exposure variable: infant and maternal death case study. In *Journal of Physics: Conference Series*, volume 1752, page 012016. IOP Publishing.

Marshall, A. W. and Olkin, I. (1990). Multivariate distributions generated from mixtures of convolution and product families. *Lecture Notes-Monograph Series*, pages 371–393.

McKenzie, E. (1985). Some simple models for discrete variate time series 1. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability*, pages 679–705.

McKenzie, E. (1988). Some arma models for dependent sequences of poisson counts. *Advances in Applied Probability*, pages 822–835.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Nastić, A. S., Ristić, M. M., and Popović, P. M. (2016). Estimation in a bivariate integer-valued autoregressive process. *Communications in Statistics-Theory and Methods*, 45(19):5660–5678.

Nelson, D. B. (1990). Arch models as diffusion approximations. *Journal of econometrics*, 45(1-2):7–38.

Nikoloulopoulos, A. K. (2013). Copula-based models for multivariate discrete response data. *Copulae in Mathematical and Quantitative Finance*, pages 231–249.

Nikoloulopoulos, A. K. (2016). Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic environmental research and risk assessment*, 30(2):493–505.

Nikoloulopoulos, A. K. and Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, 37(9):1555–1568.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Pavlopoulos, H. and Karlis, D. (2008). INAR(1) modeling of overdispersed count series with an environmental application. *Environmetrics: The official journal of the International Environmetrics Society*, 19(4):369–393.

Pechon, F., Denuit, M., and Trufin, J. (2019). Multivariate modelling of multiple guarantees in motor insurance of a household. *European Actuarial Journal*, 9(2):575–602.

Pechon, F., Denuit, M., and Trufin, J. (2021). Home and motor insurance joined at a household level using multivariate credibility. *Annals of Actuarial Science*, 15(1):82–114.

Pechon, F., Trufin, J., and Denuit, M. (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin: The Journal of the IAA*, 48(3):969–993.

Pedeli, X. and Karlis, D. (2011). A bivariate INAR(1) process with application. *Statistical modelling*, 11(4):325–349.

Pedeli, X. and Karlis, D. (2013a). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis*, 34(2):206–220.

Pedeli, X. and Karlis, D. (2013b). On estimation of the bivariate poisson INAR process. *Communications in Statistics-Simulation and Computation*, 42(3):514–533.

Popović, P. M. (2016). A bivariate INAR (1) model with different thinning parameters. *Statistical Papers*, 57(2):517–538.

Popović, P. M., Ristić, M. M., and Nastić, A. S. (2016). A geometric bivariate time series with different marginal parameters. *Statistical Papers*, 57(3):731–753.

Ristić, M. M., Bakouch, H. S., and Nastić, A. S. (2009). A new geometric first-order integer-valued autoregressive (nginar (1)) process. *Journal of Statistical Planning and Inference*, 139(7):2218–2226.

Ristić, M. M., Nastić, A. S., Jayakumar, K., and Bakouch, H. S. (2012). A bivariate INAR (1) time series model with geometric marginals. *Applied Mathematics Letters*, 25(3):481–485.

Scotto, M. G., Weiß, C. H., and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, 15(6):590–618.

Serfozo, R. F. (1972). Conditional poisson processes. *Journal of Applied Probability*, 9(2):288–302.

Shi, P. and Valdez, E. A. (2014a). Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal*, 2014(2):159–179.

Shi, P. and Valdez, E. A. (2014b). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, 55:18–29.

Sichel, H. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society: Series A (General)*, 137(1):25–34.

Silva, A., Rothstein, S. J., McNicholas, P. D., and Subedi, S. (2019). A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC bioinformatics*, 20(1):1–11.

STCHKL, H. (1982). Asymptotic efficiencies of three methods of estimation for the inverse gaussian-poisson distribution. *Biometrika*, 69(2):467–472.

Stein, G. Z. and Juritz, J. M. (1988). Linear models with an inverse gaussian poisson error distribution. *Communications in Statistics-Theory and Methods*, 17(2):557–571.

Tavaré, S. (2018). The linear birth–death process: an inferential retrospective. *Advances in Applied Probability*, 50(A):253–269.

Tzougas, G. and di Cerchiara, A. P. (2021). The multivariate mixed negative binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics*, 101:602–625.

Tzougas, G. and di Cerchiara, A. P. (2023). Bivariate mixed poisson regression models with varying dispersion. *North American Actuarial Journal*, 27(2):211–241.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(1):1–45.

Weiß, C. H. (2008a). Serial dependence and regression of poisson inarma models. *Journal of Statistical Planning and Inference*, 138(10):2975–2990.

Weiß, C. H. (2008b). Thinning operations for modeling time series of counts—a survey. *AStA Advances in Statistical Analysis*, 92(3):319–341.

Weiß, C. H. (2015). A poisson INAR(1) model with serially dependent innovations. *Metrika*, 78(7):829–851.

Weiß, C. H. (2018). *An introduction to discrete-valued time series*. John Wiley & Sons.

Xu, J., Guttorp, P., Kato-Maeda, M., and Minin, V. N. (2015). Likelihood-based inference for discretely observed birth–death-shift processes, with applications to evolution of mobile genetic elements. *Biometrics*, 71(4):1009–1021.

Zheng, H., Basawa, I. V., and Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference*, 137(1):212–229.