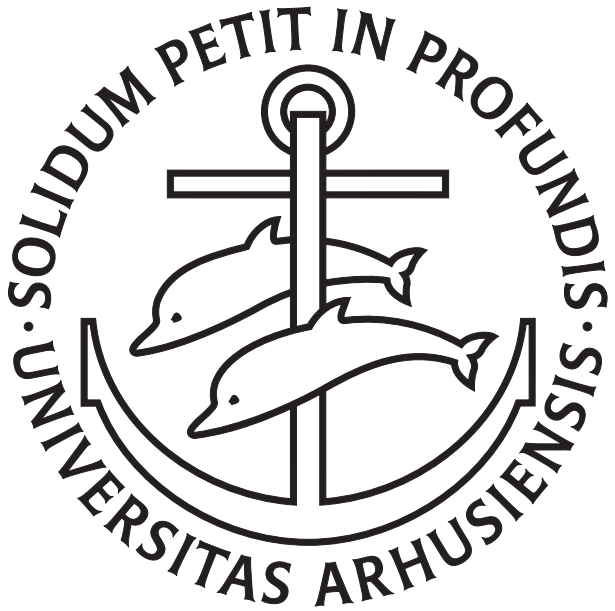
Models and Inference for Correlated Count Data



Camilla Mondrup Andreassen

PhD Dissertation
Advisor: Jens Ledet Jensen
July 2013

Department of Mathematics
Aarhus University

Contents

Accompanying Papers

Preface

This thesis presents the results of the research part of my work during my PhD studies at Department of Mathematics, Aarhus University. The topic of the thesis is, as the title suggests, models and inference for correlated count data with an emphasis on multivariate Poisson models. The thesis consists of an introductory chapter followed by three independently written papers.

I would like to take this opportunity to thank some of the people who have had an important and positive influence on my time as a PhD student.

Foremost, I would like thank my advisor Prof. Jens Ledet Jensen for his invaluable guidance. Without his many insightful comments, suggestions and corrections I would not have been able to write this thesis.

During the autumn of 2012 I had the pleasure of visiting Prof. Richard A. Davis at the Department of Statistics, Columbia University. I would like to thank him for the many hours of interesting discussions and his general interest in my work. In addition, I would like to express my gratitude for the great hospitality and the socially enjoyable environment both to him, his wife and the PhD students at the Department of Statistics.

Furthermore, I would like to thank all of my current and former fellow students, colleagues and educators at Department of Mathematics for providing an excellent workplace as well as a great social environment. In that regard, a special acknowledgement goes to Prof. Eva Vedel Jensen and CSGB[1] as well as the people who at some point during the last four years have frequented either B3.15 or B3.29. In particular, I want to thank Ólöf Thórisdóttir for being both a great colleague and a very dear friend to me during our time as office mates.

Last but not least, I want to thank my friends and family for their love and support and for taking my mind off statistics whenever I needed it.

Camilla Mondrup Andreassen
Aarhus, July 2013

[1] Centre for Stochastic Geometry and Advanced Bioimaging

Summary

In statistics, *count data* are data in which the observations can take only the non-negative integer values and are often observed in applied fields such as biology and epidemiology, for instance. The well-known univarate discrete distributions, e.g., the Poisson and the negative binomial distribution do not generalise to the multivariate case in a natural way. Therefore, how to model multivariate count data is not obvious. This thesis deals with several issues related to the analysis of correlated count data. It contains an introduction followed by three papers.

Paper A

In this paper, a certain class of the multivariate negative binomial distribution for which the distribution is defined by its probability generating function is considered. In the general case, the probability function can be expressed via the so-called α*-permanent*, which can be thought of as a generalisation of the determinant. No closed form expression for the α-permanent exists, and the probability function has therefore previously only been derived for the two-dimensional case. As a consequence hereof, inference for this distribution has been restricted to the use of composite likelihood based on one- or two-dimensional marginals. In this paper we derive the three-dimensional probability function as a sum with positive terms only and study the range of possible parameter values. The subclass of infinitely divisible distributions is considered in order to obtain more explicit results.

Paper B

Here, a multivariate mixed Poisson model is considered. The model is a generalisation of previously published Poisson mixture models since the mixing variable here arises as a function of independent and identically distributed random variables. The common distribution of the mixing variables belongs to an exponential family, a generalisation of the often considered gamma distribution. For multivariate mixture models the probability function is only rarely tractable and a composite likelihood based on the two-dimensional marginals is considered. The main result of the paper gives conditions for existence, consistency and asymptotic normality of the maximum pairwise likelihood estimate.

Paper C

The focus of this paper is on modelling bivariate time series of counts. Two versions of a Poisson-based, bivariate INGARCH model are considered. The models only differ in the construction of the bivariate, conditional Poisson distribution, since the conditional mean process has the same structure for both models. For the first model, a stability result already obtained in the literature is generalised to the case of an exponential family for the conditional distribution of the counts, and regularity conditions for strong consistency of the maximum likelihood estimate are derived. A limitation of the first model is its inability to capture negative dependence, and the second model, based on a copula approach, is therefore proposed. Stability properties for the new model are derived, and the two models are compared through a simulation study and application to a real data example. This is work in progress.

Resumé

Tælledata er data, hvor observationer kun antager de ikke-negative heltal og observeres ofte inden for anvendte videnskaber som f.eks. biologi og epidemiologi. Det er ikke helt oplagt, hvordan flerdimensionalt tælledata skal modelleres, idet flere af de velkendte endimensionale fordelinger, som f.eks. Poisson-fordelingen og den negative binomial-fordeling, ikke generaliserer naturligt til flere dimensioner. Denne afhandling beskæftiger sig med flere problemstillinger, der relaterer til analyse af korreleret tælledata. Afhandlingen består af en introduktion efterfulgt af tre artikler.

Artikel A

I denne artikel betragtes en klasse af den flerdimensionale negative binomialfordeling, der er defineret gennem sin sandsynlighedsgenererende funktion. I det generelle tilfælde kan sandsynlighedsfunktionen udtrykkes gennem den såkaldte α-permanent, der kan betragtes som en generalisering af determinanten. Der findes ikke noget lukket udtryk til beregning af α-permanenten, og sandsynlighedsfunktionen er derfor tidligere kun blevet udledt i det todimensionale tilfælde. En konsekvens heraf er, at inferens for denne fordeling er begrænset til brugen af composite likelihood-estimation baseret på en- og todimensionale marginaler. I denne artikel udledes en formel for den tredimensionale sandsynlighedsfunktion, og formlen er en sum, der udelukkende består af positive led. Derudover betragtes klassen af mulige parameterværdier, og delklassen af uendeligt delbare fordelinger betragtes med henblik på at opnå mere eksplicitte resultater.

Artikel B

I denne artikel betragtes en mixed-Poisson-model. Modellen er en generalisering af tidligere publicerede mixed-Poisson-modeller, idet mixing-variablen her opstår som en funktion af uafhængige og identisk fordelte stokastiske variable. Den fælles fordeling for mixing-variablene tilhører en eksponentiel familie, en generalisering af de ofte betragtede gamma fordelinger. For flerdimensionale mixturfordelinger er den fælles sandsynlighedsfunktion ofte uhåndterbar, og derfor betragtes composite likelihood-estimation baseret på de todimensionale marginaler. Hovedresultatet i

denne artikel giver betingelser for eksistens, konsistens og asymptotisk normalitet for maksimum composite likelihood-estimatet.

Artikel C

I dette manuskript er fokuspunktet modellering af todimensionale tidsrækker af tællevariable. Vi betragter to versioner af en todimensional INGARCH-model. Forskellen på de to modeller ligger kun i konstruktionen af den todimensionelle betingede Poisson-fordeling, idet den betingede middelværdiproces har samme struktur for begge modeller. For den første model betragter vi et stabilitetsresultat, tidligere vist i litteraturen, og viser, at dette generaliserer til det tilfælde, hvor den todimensionelle betingede fordeling tilhører en eksponentiel familie. Derudover udledes betingelser, der sikrer konsistens af maksimum likelihood-estimatet. Ud fra den måde hvorpå Poisson-fordelingen i denne model er konstrueret, giver modellen kun mulighed for at modellere positiv afhængighed mellem de to tidsrækker. Vi foreslår derfor, at modellere afhængigheden vha. copulaer. Vi udleder betingelser, der sikrer stabilitet af den nye model, og til slut sammenlignes de to modeller gennem et simuleringsstudium samt anvendelse på data. Arbejdet præsenteret i dette manuskript er igangværende forskning.

Introduction

The purpose of the present chapter is to give an introduction to the subjects covered in the thesis and describe the main results obtained.

1.1 Multivariate count data

In statistics, *count data*, or discrete data, are data in which observations can take only non-negative integer values and where these arise from counting. Simple examples of count data are the number of radioactive decays over a given time interval, number of mutations on a DNA strand per unit length or the number of days a machine works before it breaks down. When counts are treated as independent random variables, the Poisson, binomial and negative binomial distributions are commonly used to model the data.

This thesis deals with models and inference for correlated count data which are frequently encountered in applied fields, e.g. epidemiology, biology, marketing, criminology, accident analysis etc. A simple example of correlated count data is when data are collected in a field, and where the 'living conditions' vary over the field. This is the case in the following two data sets. The data of Choo and Walker (2008) reports the occurrence of testis cancer in 19 municipalities in the county of Frederiksborg, Denmark, and the data of Beall (1939) reports the number of Colorado potato beetles in each of 2304 units of a single field. For data of this type, counts at nearby positions are, due to this underlying variation, expected to be more similar than counts at distant positions and they are therefore likely to be positively correlated.

An example of negatively correlated count data can be found in Aitchison and Ho (1989). The data originates from a study of the relative effectiveness of three different air samplers used for detecting pathogenic bacteria in sterile rooms. The resulting data consists of triplets of bacterial colony counts from the samplers in 50 different sterile rooms. The data can therefore be described by a trivariate discrete distribution. Aitchison and Ho (1989) model the data by use of a Poisson log-normal mixture model and maximum likelihood estimation yields a significantly negative correlation between the counts. This indicates that the samplers appear to have been competing for the capture of the bacteria. In general, simple examples of

negative correlated count data are data that can be modelled by the multinomial distribution, e.g., the number of votes the different candidates receive at an election where the number of voters is fixed - an increase in the number of votes on one candidate requires a decrease in the number of votes on another candidate.

As mentioned above, one-dimensional count data are commonly modelled by the Poisson, the binomial or the negative binomial distribution. These distributions are well studied and estimation of the parameters can easily be performed by use of maximum likelihood estimation. The situation is not that simple for correlated count data. The following sections review methods to overcome some of the problems that arise when multivariate data is considered.

1.2 Construction of multivariate distributions

It is well known that the Gaussian distribution generalises in a natural way from one dimension to two or higher dimensions. The probability density function (pdf) of a Gaussian random variable X is given as

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\tfrac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R},$$

with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The equivalent for n dimensions is given as

$$f_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad x \in \mathbb{R}^n,$$

with $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n) \in \mathbb{R}^n$ and $\Sigma = \{\sigma_{ij}\}_{i,j=1,\ldots,n}$ being a symmetric positive definite matrix. It follows that $\mathbb{E}X_i = \mu_i$ and $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$ for $i, j = 1,\ldots,n$. Furthermore, the marginal distributions are normal with means μ_i and variance σ_{ii}, $i = 1,\ldots,n$.

Another well-known example where the one-dimensional case generalises easily to higher dimensions is multidimensional contingeny tables (Lauritzen, 2002). Let us first recall the multinomial distribution. Consider n identical trials with k possible outcomes. Let X_i be the number of times that outcome i occurs in the n trials, $i = 1,\ldots,k$. It is assumed that the probability of outcome i is the same for all trials and this probablity is denoted p_i. The probability mass function (pmf) of the multinomial distribution is then given as

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!}{x_1!\ldots x_k!}\prod_{i=1}^{k} p_i^{x_k}.$$

This is the usual characterisation of the multinomial distribution. One could also consider the n objects as being classified according to a criterion with k different levels. This is the case for a one-dimensional contingency table. An m-dimensional contingency table is defined the following way. For ease of notation we describe the two-dimensional case but the principle is the same for higher dimensions. Let n objects be classified according to two criteria, A and B, having levels $A_1,\ldots,A_r$ and $B_1,\ldots,B_s$. This gives rise to an $r \times s$ table on the form $\{n_{ij}\}$ with n_{ij} being the number of objects classified according to the levels A_i and B_j. If the probability

of being classified as level A_i and B_j is p_{ij}, then the statistical model of the data is a multinomial distribution with rs possible outcomes, having n as the number of trials and p_{ij} as the probability of outcome (A_i, B_j). A well-known class of models for contingency tables are the Hierarchical models with the subclasses decomposable models, graphical models and Markov random fields (Darroch et al., 1980).

Unfortunately, not all well-known (one-dimensional) distributions generalise this easily to higher dimensions. One of the challenges is the difficulty in finding a distribution that covers the entire range of possible dependences. Construction of multivariate distributions is therefore a classical and ongoing field of research in statistics. In this section we recall a few methods for constructing multivariate distributions considered in this thesis. An extensive overview on the topic can be found in Alzaga and Déniz (2008), and comprehensive studies of bivariate and multivariate discrete distributions in general can be found in Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997), respectively.

1.2.1 Trivariate reduction

Trivariate reduction, also known as the *variables in common* method, is a well-known and easy-to-apply technique for constructing dependent variables (Mardia, 1970). The method applies in the continuous as well as the discrete case. When applying the method to construct a pair of dependent variables, Y_1 and Y_2, the starting point is three mutually independent random variables, X_1, X_2 and X_3. The variables are then connected through functions, f_1 and f_2, in the following way.

$$Y_1 = f_1(X_1, X_3) \quad \text{and} \quad Y_2 = f_2(X_2, X_3).$$

Often the functions f_1 and f_2 are simple functions, e.g. summation of the variables. Johnson et al. (1997) defines the bivariate Poisson distribution as the joint distribution of the random variables $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$, where X_1, X_2 and X_3 are mutually independent Poisson random variables (this method is considered in Paper C). A limitation of this method is that it only covers the case with positive correlation since $\mathrm{Cov}(Y_1, Y_2) = \mathrm{Var}(X_3) \geq 0$. The multivariate gamma distribution considered in Choo and Walker (2008) can be considered as a generalisation of this method, see section 1.2.2.

A simple way to introduce negative correlation for continuous random variables is the following way. Let U be a uniform random variable on the interval $(0,1)$ and define the random variables $Y_1 = F_\lambda^{-1}(U)$ and $Y_2 = F_\lambda^{-1}(1-U)$ where F_λ is a continuous and strictly increasing cumulative distribution function (cdf) (e.g. the cdf of the exponential distribution). Since $1 - U$ is also uniform on the interval $(0,1)$ it now follows that $Y_i \sim F_\lambda$, $i = 1, 2$, and the correlation between Y_1 and Y_2 is negative due to Chebyshev's other inequality (Fink and Jodeit Jr, 1984). Shin and Pasupathy (2007, 2010) use this idea and present an algorithm, the Trivariate Reduction Extension algorithm, for generating bivariate negatively correlated Poisson variables. Let $\rho \in (-1, 1)$, $\lambda \geq \lambda' > 0$ and $k = \lambda'/\lambda$. The principle is to construct dependent Poisson random variables Y_1 and Y_2, having correlation ρ and means λ

and λ' respectively, the following way. Let

$$\begin{aligned} Y_1 &= X_1 + F^{-1}_{\lambda^*}(U), \quad Y_2 = X_2 + F^{-1}_{k\lambda^*}(U) && \text{if } \rho > 0, \\ Y_1 &= X_1 + F^{-1}_{\lambda^*}(U), \quad Y_2 = X_2 + F^{-1}_{k\lambda^*}(1-U) && \text{if } \rho < 0, \end{aligned}$$

where U is uniform on the interval $(0,1)$, F_λ is the Poisson cdf with mean λ and $F^{-1}_\lambda(u) = \inf\{x : F_\lambda(x) > u\}$. It follows that $X_1 \sim \text{Po}(\lambda - \lambda^*)$ and $X_2 \sim \text{Po}(\lambda' - k\lambda^*)$. The main hurdle is to solve for λ^* such that Y_1 and Y_2 attain the target means, λ and λ', and the target correlation, ρ. Shin and Pasupathy (2010) present a fast numerical procedure to identify λ^*.

1.2.2 Mixture models

Mixture distributions are often seen in the literature. When a population contains two or more homogeneous subpopulations, with two subpopulations being the simplest case, mixture models are a natural choice. When sampling from a population consisting of two subpopulations denoted A and B we sample from A with probability p and from B with probability $1-p$ for some $p \in (0,1)$. The cdf for the population is then given as $pF_A + (1-p)F_B$ where F_A and F_B are the cdf's of A and B, respectively. In general, if the number of subpopulations or mixing distributions is finite the cdf can be expressed as a weighted sum of the cdf's of the subpopulations (Lindsay, 1995).

In this section we consider the case where the number is uncountable, since the model considered in Paper B is of this type. A distribution of this type is the distribution that results from assuming that a random variable is distributed according to some parameterised distribution and that the parameters of that distribution are random variables themselves. Mathematically formulated, let the random variable X have density (or pmf) f_λ with λ having density g_θ. The resulting distribution of X is then given as

$$p_X(x;\theta) = \int f_\lambda(x;\theta) g_\theta(\lambda)\, d\lambda. \tag{1.1}$$

A formula of the same type applies if some or all of the variables are vectors.

One reason for using mixture models is the possibility of modelling overdispersion, which is often found in correlated count data. If f_λ denotes the pmf of a Poisson distribution, (1.1) represents a mixed Poisson distribution (Grandell, 1997). It follows that $\mathbb{E}\, X = \mathbb{E}\,\lambda$ and $\text{Var}(X) = \mathbb{E}\,\lambda + \text{Var}(\lambda) \geq \mathbb{E}\,\lambda$. Hence, the mixed Poisson distribution is overdispersed relative to the Poisson distribution.

Some well-known examples of count mixtures are the negative binomial distribution and the beta-binomial distribution. The negative binomial distribution arises as a continuous mixture of a Poisson distribution $\text{Po}(\lambda)$ where the parameter λ is a random variable with a gamma distribution $\text{Ga}(r, (1-p)/p)$ with $0 < r < \infty$ and $0 < p < 1$. The beta-binomial distribution arises as a mixture of a binomial distribution $\text{bi}(n,p)$ where p is a random variable distributed according to a beta distribution $\text{Be}(\alpha,\beta)$ with $0 < \alpha, \beta < \infty$. A renowned example of a mixture of continuous variables is the generalised hyperbolic distribution, which arises as a mixture of the normal distribution and the generalised inverse Gaussian distribution (Barndorff-Nielsen, 1978).

From the above it follows that mixture distributions give a simple way of constructing a multivariate distribution. For instance, consider the model of Choo and Walker (2008). The model proposed in that paper is a Poisson-gamma model for investigating spatial variations of disease. Other Poisson-gamma models have been proposed (Clayton and Kaldor (1987) and Tsutakawa (1988) among others) but these fail to model the spatial correlation. The observations considered are disease counts in different areas and these are modelled as conditionally independent Poisson random variables, i.e. $Y_i|\theta_i \sim \text{Po}(\mu_i)$, $i = 1,\ldots,N$, with $\mu_i = E_i\theta_i$ where E_i is a positive finite constant. The variable θ_i is interpreted as the relative risk in area i. To construct a multivariate distribution for $\theta_1,\ldots,\theta_N$ independent gamma random variables $R_{i,j} \sim \text{Ga}(a,a)$ are introduced. They can be thought of as controlling the dependence between a neighbouring pair of areas i and j. With A_i being the set of pairs of neighbours involving area i, θ_i is defined as $\theta_i = n_i^{-1}\sum_{(i,j)\in A_i} R_{i,j}$ with n_i being the number of elements in A_i. The resulting distribution is multivariate with gamma marginals having mean one and variance $(n_i a)^{-1}$ and $\text{Cov}(\theta_i,\theta_{i'}) = (n_i n_{i'} a)^{-1}$, $i \neq i'$ (see also Johnson and Kotz (1972) for construction of multivariate gamma distributions by sums of independent gamma variables). Therefore, the resulting model for the counts is a multivariate mixed Poisson model with $\text{Cov}(Y_i, Y_{i'}) = \text{Cov}(\theta_i,\theta_{i'}) > 0$. Another example of a multivariate mixture model is the model of Henderson and Shimakura (2003), see the following section.

1.2.3 Laplace transform or probability generating function

Some classes of multivariate distributions (e.g. gamma- and negative binomial-type distributions) are defined solely through specification of the Laplace transform or the probability generating function (pgf) and does not in general have a closed form expression for the resulting probability distribution. Consider the following example of the construction of a multivariate gamma distribution. Let $Y_1,\ldots,Y_q$ be independent p-dimensional Gaussian random vectors with standard marginals and a common correlation matrix C $(p\times p)$. Define $Z_k = \frac{1}{q}\sum_{j=1}^{q} Y_{jk}^2$ for $k = 1,\ldots,p$. The resulting distribution of the vector $Z = (Z_1,\ldots,Z_p)^T$ is a multivariate gamma distribution with marginal gamma distributions, $\text{Ga}(q/2, q/2)$, and Laplace transform

$$\mathbb{E}[e^{-u^T Z}] = \left|I + \tfrac{2}{q}CU\right|^{-q/2}, \tag{1.2}$$

with $U = \text{diag}(u)$, $|\cdot|$ being the determinant and I is the $p\times p$ identity matrix (Krishnamoorthy and Parthasarathy, 1951).

Before continuing, we introduce the so-called *α-permanent* of a real square matrix. Let $A = \{a_{ij}\}$ be a real $m\times m$ matrix and $\alpha \in \mathbb{R}$. The α-permanent of A is defined as

$$\text{per}_\alpha(A) = \sum_{\sigma\in\mathcal{S}_m} \alpha^{c(\sigma)} a_{1,\sigma(1)} a_{2,\sigma(2)} \ldots a_{m,\sigma(m)},$$

where $\mathcal{S}_m$ is the set of all permutations of $1,\ldots,m$ and $c(\sigma)$ denotes the number of cycles in the permutation σ. We notice that when $\alpha = -1$ this is simply the determinant of A. As is the case with the determinant, the α-permanents are in general computationally complex.

Vere-Jones (1997) studied the α-permanents and their application to the multivariate negative binomial distribution, among other distributions. Let N be an m-dimensional discrete random vector with pgf

$$\mathbb{E}\prod_{i=1}^{m} z_i^{N_i} = |I + (I - Z)A|^{-\alpha}, \tag{1.3}$$

where $\alpha > 0$ and A is a real $m \times m$ matrix. The one-dimensional marginals of N can be shown to be negative binomial and we therefore refer to this distribution as the multivariate negative binomial distribution and denote it by $\mathrm{NB}_m(\alpha, A)$, see Paper A. With $Q = A(I + A)^{-1}$ the resulting pmf is given as

$$\mathbb{P}(N = n) = |I - Q|^{\alpha} \frac{\mathrm{per}_{\alpha}(Q[n])}{\prod_{i=1}^{m} n_i!}, \tag{1.4}$$

where $Q[n]$ is obtained from Q by repeating index i n_i times (Vere-Jones, 1997).

The question of existence of the model defined by (1.3) is treated in detail in Griffiths and Milne (1987), Vere-Jones (1997), Shirai (2007) and Paper A. For a general m there does not seem to exist easily verifiable necessary and sufficient conditions. However, there do exist simple sufficient conditions, two of which is given below (see the aforementioned references).

1. A is a covariance matrix and $\alpha = \frac{k}{2}$, $k = 1, 2 \ldots$.
2. Q has non-negative entries and the eigenvalues of Q is bounded by one.

Loosely speaking, the conditions ensure that $|I - Q| > 0$ and $\mathrm{per}_{\alpha}(Q[n]) \geq 0$ for all $n \in \mathbb{N}_0^m$ which is necessary for (1.4) to be a true probability function.

Møller and Rubak (2010) considered the so-called *α-permanental random field* (α-prf) which plays an important role in the study of α-permanental point processes (see for instance McCullagh and Møller (2006) and the references therein). The model can, however, be considered as a special case of the multivariate negative binomial distribution. The random vector $(N_1, \ldots, N_m)$ is an α-prf with parameter (α, C) if the pgf is given as $\mathbb{E}\prod_{i=1}^{m} z_i^{N_i} = |I + \alpha(I - Z)C|^{-1/\alpha}$ with α a positive number and C a real $m \times m$ matrix. This corresponds to the class $\mathrm{NB}_m(1/\alpha, \alpha C)$.

In some parts of the parameter space the distribution resulting from (1.3) can in fact be obtained as a mixture distribution where the mixing distribution is multivariate gamma. This is the case in Henderson and Shimakura (2003) where a Poisson-gamma model for longitudinal count data is presented. The multivariate gamma distribution considered is defined through its Laplace transform, a generalised version of (1.2). Mathematically, with $N_1, \ldots, N_p$ denoting the event counts and $Z_1, \ldots, Z_p$ the corresponding frailties, the counts are conditionally independent Poisson random variables

$$N_j \,|\, Z_j \sim \mathrm{Po}(Z_j e^{x_j^T \beta}), \quad j = 1, \ldots, p,$$

with x_j being a fixed covariate vector and β an unknown regression coefficient. The Laplace transform of the joint distribution of $Z_1, \ldots, Z_p$ is given as $\mathbb{E}\, e^{-u^T Z} = |I + \xi C U|^{-1/\xi}$ with $\xi > 0$ and C being a $p \times p$ matrix with entries $C_{jk} = \rho^{|j-k|}$,

$0 \leq \rho \leq 1$. Hence, the univariate marginals are gamma with mean one, variance ξ and $\text{Cor}(Z_j, Z_k) = \rho^{|j-k|}$. The pgf of $(N_1, \ldots, N_p)$ can then be found to be

$$\mathbb{E}\prod_{j=1}^{p} z_j^{N_j} = \mathbb{E}\Big[\exp\Big\{-\sum_{j=1}^{p}(1-z_j)u_j Y_j\Big\}\Big] = |I + \xi(I-Z)UC|^{-1/\xi},$$

with $u_j = e^{x_j^T \beta}$. This is recognised as the pgf of $\text{NB}_m(1/\xi, \xi UC)$. See also Chatelain et al. (2009) for an example.

For later reference, we note that an enhanced version of the model of Henderson and Shimakura (2003) is proposed in Fiocco et al. (2009). The setup is the same but Fiocco et al. (2009) present a new gamma process which possesses the same moments of interest as the one considered in Henderson and Shimakura (2003). An advantage of the new process is that the finite dimensional marginals have distributions that are computationally more stable and it is possible to simulate from the entire parameter space. The distribution is construced by use of a renewal process and the fact that a gamma distribution is infinite divisible. We refer to Fiocco et al. (2009) for further details.

1.2.4 Copulas

During the recent years copulas have become very popular for modelling multivariate non-normal data. Loosely speaking, copulas are functions that couple marginal distribution functions. That is, a cdf can be written in terms of (one-dimensional) marginal distribution functions and a copula, where the marginal distribution functions describe the distributions of the marginals, and the copula describes the dependence between the marginals. The idea behind the copulas can be described as follows. Let $X = (X_1, \ldots, X_d)$ be a random vector with continuous marginals. Let $F_i(x) = \mathbb{P}(X_i \leq x)$ be the associated cdfs. We note that they are continuous functions due to the continuous marginals of X. Then, the vector $U = (U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$ has uniform marginals and the copula C of X is defined as the joint cdf of U, i.e. $C(u_1, \ldots, u_d) = \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d)$. The copula C then contains all the information about the dependence between the variables and the F_is contain all the information about the marginal distributions. This leads to the following definition (Joe, 1997; Nelsen, 2006).

Definition 1.1. *$C : [0,1]^d \to [0,1]$ is a d-dimensional copula if C is a joint cdf of a d-dimensional random vector on the unit cube $[0,1]^d$ with uniform marginals.*

Copulas have one very important property, which lays the theoretical foundation for their use. It is given in Sklar's theorem below.

Theorem 1.2 (Sklar, 1959). *Let H be the joint cdf of $(X_1, \ldots, X_d)$ with marginal cdfs $F_i(x) = \mathbb{P}(X_i \leq x)$. Then there exists a copula C such that for all $x_1, \ldots, x_d \in \mathbb{R}$*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{1.5}$$

If $F_1, \ldots, F_d$ are continuous, then C is unique. Otherwise C is uniquely determined on $\text{Ran}(F_1) \times \cdots \times \text{Ran}(F_d)$, i.e. the Cartesian product of the ranges of the marginal cdfs.

Conversely, if $F_1,\dots,F_d$ are cdfs and C is a copula, then H defined by (1.5) *is a cdf with marginals $F_1,\dots,F_d$.*

The second part of Sklar's theorem states that given a marginal distribution function and a copula one can construct a multivariate cdf. Hence, using a copula to model the joint distribution of a set of random variables, makes it possible to specify the marginal distributions in a more flexible way since one does not have to specify the entire model at once. In particular, the marginal distributions can be selected separately, which gives the possibility of having different types of marginal distributions, e.g., a valid distribution can be constructed by combining a normal distribution with a gamma distribution by use of the Frank copula (section C.4.1). A review on copulas for count data can be found in Genest and Neslehova (2007) and examples of modelling multivariate count data based on various copulas can be found in Nikoloulopoulos and Karlis (2009). We also refer to Trivedi and Zimmer (2005).

When modelling multivariate data by use of copulas, the derivation of the joint density is easy for the continuous case as it can be found through partial derivatives of the copula cdf. This is not possible in the discrete case where instead the pmf can be found in the following way.

Proposition 1.3. *Consider a discrete integer-valued random vector $(Y_1,\dots,Y_m)$ with marginals $F_1,\dots,F_m$ and joint cdf given by the copula representation $H(y_1,\dots,y_m) = C(F_1(y_1),\dots,F_m(y_m))$. Let $x = (x_1,\dots,x_m)$ with x_k being equal to either y_k or $y_k - 1$, $k = 1\dots,m$. The joint pmf of $(Y_1,\dots,Y_m)$ is then given by*

$$h(y_1,\dots,y_m) = \sum_x (-1)^{n(x)} C(F_1(x_1),\dots,F_m(x_m)), \tag{1.6}$$

with $n(x)$ being the number of x_k's equal to $y_k - 1$.

From the proposition is follows that in order to calculate the pmf the copula has to be evaluated repeatedly. Therefore, in order to model multivariate count data by use of copulas one must use a copula with a computationally feasible form of the corresponding cdf.

The family of copulas is very extensive. In this thesis we restrict attention to the subclass of Archimedean copulas (Paper C). The two main reasons for considering Archimedean copulas is that they (1) allow modelling of dependence in arbitrarily high dimensions with only one parameter and (2) the most common Archimedean copulas (e.g., the Frank and the Clayton copulas) have a closed form expression for the cdf (contrary to for example the multivariate elliptical copulas, e.g., the Gaussian and the Student-t copula). The Archimedean copulas are defined as follows.

Definition 1.4. *Let $\psi : [0,1] \to [0,\infty]$ be a continuous and strictly decreasing convex function with $\psi(1) = 0$. The Archimedean copula is defined as*

$$C(u_1,u_2) = \psi^{[-1]}(\psi(u_1) + \psi(u_2)),$$

with $\psi^{[-1]}$ *denoting the pseudo-inverse of* ψ *with domain* $[0,\infty]$ *defined by*

$$\psi^{[-1]}(t) = \begin{cases} \psi^{-1}(t), & 0 \le t \le \psi(0), \\ 0, & \psi(0) \le t \le \infty. \end{cases}$$

For examples of Archimedean copulas we refer to section C.4.1.

The topic of Paper C is modelling of bivariate time series of counts, and in particular the use of copulas for modelling the dependence between the two time series at a given point in time. Heinen and Rengifo (2007) propose the Multivariate Autoregressive Conditional Double Poisson model for modelling multivariate time series of counts, a multivariate extension to the univariate model developed in Heinen (2003). The marginal distributions are modelled by the double Poisson distribution (Efron, 1986), which allows for under- and overdispersion (contrary to the regular Poisson distribution), and the dependence between the observed counts at a given point in time is modelled by a Gaussian copula. In order to be able to apply the density of a Gaussian copula Heinen and Rengifo (2007) use the continuous extension argument of Denuit and Lambert (2005) and create a continuous version of the observed variable. The new variable is constructed by adding an independent continuous random variable, with values in $(0,1)$ and a strictly increasing cdf (e.g., the uniform distribution), to the observed variable. Applying the cdf of the new variable results in a variable being uniform on $(0,1)$ and the Gaussian copula density can be applied. We refer to Heinen and Rengifo (2007) for further details. For a review on copula models for economic time series (the continuous case) we refer to Patton (2012).

1.3 Estimation

Given data that can be described by a parameterised model, maximum likelihood estimation (MLE) is often a desirable method of estimation of the parameters. In order to use this method an explicit expression for the joint density of the data is needed. Unfortunately, it is not always available (see for instance the Papers A and B).

When a closed form expression for the joint density is not known composite likelihood methods can be very useful. Consider an m-dimensional random vector Y, with probability density function parameterised by $\theta \in \Theta \subseteq \mathbb{R}^p$. Consider a set of conditional or marginal events with associated likelihoods, $\mathcal{L}_k$, $k = 1,\dots,K$, that can be written in closed form. The composite likelihood (Lindsay, 1988) combines these events as

$$\mathcal{L}_C(\theta; y) = \prod_{k=1}^{K} \mathcal{L}_k(y;\theta)^{\omega_i},$$

where ω_k are non-negative weights to be chosen. Often, the weights are chosen such that the composite likelihood coincides with the full likelihood in the case of independence, see for instance Le Cessie and Van Houwelingen (1994) and Henderson and Shimakura (2003). As in maximum likelihood estimation, the maximum composite likelihood estimator is found by maximising the composite likelihood or equivalently the log-composite likelihood. The composite likelihood

is a product of valid likelihoods and therefore approximately unbiased estimators with the usual asymptotic properties can be obtained under regularity conditions (Varin et al., 2011).

A simple way of constructing a composite likelihood is to consider the product of lower dimensional marginal densities, when closed form expressions for these are known, with one-dimensional marginals as the simplest choice (Cox and Reid, 2004; Varin, 2008). These composite likelihoods are sometimes referred to as composite marginal likelihoods (Varin, 2008). Using one-dimensional marginals corresponds to assuming independence between the observations and an obvious limitation of this approach is the lack of possibility to estimate parameters associated with dependence. Hence, when the observations are dependent a composite likelihood based on the two- or three-dimensional marginals is often desireable. Composite likelihoods based on two-dimensional marginals are often referred to as pairwise likelihoods. Another well-known type of composite likelihoods is the pseudolikelihood proposed in Besag (1974), a composite likelihood constructed from marginal conditional densities. An extensive overview of the theory and application of composite likelihood is given in Varin et al. (2011).

The advantages of composite marginal likelihoods in practical applications have been treated several places in the literature. For the model considered in Henderson and Shimakura (2003) a closed form expression for the full likelihood can be found by differentiation of the Laplace transform, but the number of terms to be calculated increases drastically with the number of observations. Therefore the full likelihood is intractable and pairwise likelihood as estimation method is proposed. Other applications of pairwise likelihood estimation can be found in for example Chatelain et al. (2009) and Rubak et al. (2010).

One purpose of the use of composite likelihoods is to reduce the complexity of the computations. The complexity can be further reduced by considering a two-stage estimation procedure, e.g. the one proposed in Fiocco et al. (2009). Here, the marginal parameters are first estimated by use of a composite likelihood based on one-dimensional marginals and, secondly, the correlation parameter is estimated by use of a pairwise composite likelihood.

A two-stage procedure can also be useful when MLE is possible although computationally very complex. This can for instance be the case when copulas are used for modelling multivariate count data cf. the comment after Proposition 1.3. The number of terms in (1.6) is 2^m and therefore the computational complexity increases drastically with the dimension in the case of MLE. In such a situation a two-stage estimation procedure can be useful. In the first stage the parameters of the univariate marginals are estimated using MLE and in the second stage the dependence parameters are estimated with the univariate parameters held fixed at the values obtained in the first stage. A study of the asymptotic efficiency of this procedure for copula-based models can be found in Joe (2005).

1.4 Setups and main results of accompanying papers

1.4.1 Paper A

The model considered in this paper is a particular generalisation of the negative binomial distribution to the multivariate case, which is obtained through a specification of the pgf. The probability function derived from the pgf can be expressed through the α-permanent (Vere-Jones, 1997). No simple explicit form for the probability function exists and therefore no explicit form for the probability function has been derived for the general case. An expression for the two-dimensional probability function has been derived several times in the literature using different approaches (Edwards and Gurland (1961), Henderson and Shimakura (2003), Griffiths and Milne (1987) and Rubak et al. (2010)). Inference has therefore been restricted to the use of composite likelihood based on the one- or two-dimensional marginals. The aim of this paper is to derive the three-dimensional probability function and study its properties.

The model is defined as follows. Let $N = (N_1, \ldots, N_m)$ be an m-dimensional positive discrete random vector, with pgf of the form

$$\varphi(z) = \mathbb{E} \prod_{i=1}^{m} z_i^{N_i} = |I + (I - Z)A|^{-\alpha}, \tag{1.7}$$

with $|\cdot|$ the determinant, $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$, $Z = \operatorname{diag}(z_1, \ldots, z_m)$, I the $m \times m$ identity matrix, $\alpha > 0$ and $A = \{a_{ij}\}$ a real $m \times m$ matrix. The pgf can be rewritten in terms of $Q = A(I + A)^{-1}$ as $\varphi(z) = (|I - Q|/|I - ZQ|)^{\alpha}$.

Proposition A.1 gives necessary and sufficient conditions (in terms of Q) for existence of the model as well as an explicit formula for the probability function in the two-dimensional case. It states that the distribution exists if and only if the diagonal entries are positive and bounded by one, the off-diagonal entries are of the same sign (or one should be equal to zero) and the determinant $|I - Q|$ is strictly positive. Even though these results are known we give a simple and self contained proof, which is based on expansion of the pgf. The reason for including the proof is its usefulness when deriving the results in Proposition A.2, the main result of the paper.

Proposition A.2 gives necessary and sufficient conditions for the existence of the distribution for all $\alpha > 0$ in the three-dimensional case, again in terms of Q. We notice that the condition 'for all $\alpha > 0$' corresponds to infinite divisility for this model. For a fixed α there does not seem to exist simple necessary and sufficient conditions on A for existence of the distribution. However, with a requirement of existence for all $\alpha > 0$ there is a simple characterisation that generalise the conditions from the two-dimensional case. The proposition furthermore gives an explicit form for the three-dimensional probability function and an illustration of it is given in Figure A.1.

In data applications, the parameterisation in terms of A is of interest since the moments of the distribution can be found as simple functions of α and A, see (A.2). The conditions in Proposition A.1, which treats the two-dimensional case, are given in terms of Q but can easily be rewritten in terms of A. Unfortunately, this is not the case in the three-dimensional case, at least not for a general A matrix. We therefore

restrict to the case of symmetric A matrices. This is a true restriction since not all distributions in the three-dimensional case can be represented by a symmetric matrix, contrary to the two-dimensional case. However, restricting to symmetric A matrices only have an influence on the moments of order three or higher, see (A.2). Proposition A.3 gives necessary and sufficient conditions on A for (1.7) to be a pgf of a positive, discrete, infinite divisible distribution. The proposition is followed by Corollary A.4 which gives conditions for existence of a special symmetric A matrix to be used in the data application.

Finally, the use of the three-dimensional distribution is illustrated by fitting the three-dimensional volcano data of Chatelain et al. (2009). The data consists of three images of the same scenario obtained before and after an eruption: a reference image (image 2) of the Nyiragongo volcano in Congo before an eruption and two secondary images (images 1 and 3) of the same scene acquired after the eruption. The data furthermore contain a binary image indicating the pixels of the image, which have been affected by the eruption. The purpose of the analysis is to construct a change detector, i.e. an indicator of change for each pixels, based on the correlation between the pixels of the three images. For each pixel three number of photons are observed corresponding to the three images. These will be denoted N_1, N_2 and N_3 with N_2 corresponding to the observation from the reference image. Image 1 and 3 is obtained after the eruption and it is therefore assumed that $r \leq \mathrm{Cor}(N_1, N_3)$ under the assumption that $r = \mathrm{Cor}(N_1, N_2) = \mathrm{Cor}(N_2, N_3)$. For each pixel the value of r is estimated based on an $n \times n$ window centered at the pixel and if the estimated value is below some threshold value the pixel is classified as having a change from before the eruption to after the eruption.

Appendix A.3 contains supplementary material not contained in the submitted version of the paper. Here the results are applied to the α-permanental random field (Møller and Rubak, 2010). We notice that Corollary A.7 gives a simple characterisation of infinite divisibility in the case of a symmetric matrix with identical diagonal entries. Furthermore, an additional data set it fitted by the three-dimensional distribution.

1.4.2 Paper B

In this paper, we consider a multivariate mixed Poisson model where the mixing variable arise from a function of independent and identically distributed random variables whose common distribution belongs to an exponential family. The model is a generalisation of the Poisson-gamma model of Choo and Walker (2008). A challenge of this model is that the full likelihood is only rarely tractable making maximum likelihood estimation complicated and often impossible. The aim of this paper is to consider composite likelihood estimation based on the two-dimensional marginals and to show consistency and asymptotic normality of the resulting estimator. We notice that Choo and Walker (2008) consider a Bayesian approach to the estimation within the model.

The model is defined as described in the following. Consider n areas (subsets) of $\mathbb{R}^2$ and let $\sim$ be a symmetric and reflexive neighbourhood relation, e.g. two areas are neighbours if they share a common border. Define the neighbourhood of area i as $A_i = \{j \in \{1, \ldots, n\} : i \sim j\}$ with $m_i = |A_i|$ denoting the number of neighbours of

area i. Let Y_i be the observed count in area i and X_i an associated latent variable with $X = \{X_i, i = 1,\ldots,n\}$ and $X_L = \{X_i, i \in L\}$ for a subset $L \subseteq \{1,\ldots,n\}$. The counts are modelled as conditionally independent Poisson random variables

$$Y_i \,|\, X \sim \text{Po}(\beta_i \mu(X_{A_i})), \quad i = 1,\ldots,n,$$

where $\beta_i \in \mathcal{B} \subseteq \mathbb{R}_+$ is a known covariate and $\mu(X_{A_i})$ is a positive function of m_i variables. The latent variables are assumed to be independent and identically distributed random variables with a common probability density function parameterised by $\theta \in \Theta \subseteq \mathbb{R}^d$, belonging to an exponential family. That is, the density is of the form $f_\theta(x) = a(\theta)b(x)\exp\{\varphi(\theta)\cdot t(x)\}$, where $\varphi : \Theta \to \mathbb{R}^k$ and $t : \mathbb{R} \to \mathbb{R}^k$ are known functions. The functions $\varphi(\theta)$ and μ are chosen such that $\mathbb{E}[\mu(X_{A_i})]$ is independent of θ (and typically equal to 1).

Due to the structure of the model we consider a pairwise likelihood of the form

$$l_n^2(\theta) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*} \log p_{ij}(Y_i, Y_j; \theta),$$

where p_{ij} is the bivariate probability function of a neighbouring pair Y_i and Y_j and $A_i^* = A_i \setminus \{i\}$. Theorem B.1, the main result of the paper, shows that under a set of regularity conditions there exists a local maximum of l_n^2 that is consistent and asymptotically normal. Furthermore, the result holds for any consistent solution of the likelihood equation. The theorem follows from a set of lemmas according to Jensen (2011a,b) and from the Cramér-Wold Theorem (Cramér and Wold, 1936) it follows that it suffices to show the result for the case $d = 1$. One of the regularity conditions, (B3), is used for obtaining lower bounds on the moments of the first and the second order derivatives of the log-pairwise likelihood function. Sufficient conditions for this assumption to be fulfilled is given in Proposition B.6.

To illustrate the main result a simulation study is performed. For simplicity, a setup in $\mathbb{R}$ is considered, i.e. instead of subsets of $\mathbb{R}^2$ one can think of consecutive intervals of the real line and, hence, $A_i = \{i-1, i, i+1\}$. The latent variables are modelled as independent gamma random variables with mean one and variance θ^{-1} and $\mu(X_{A_i}) = X_{i-1} + X_i + X_{i+1}$. With this setup it is possible to derive a closed form expression of the bivariate probability function of interest and composite likelihood estimation is therefore possible. Finally, the model is fitted to a data set on the occurrence of testis cancer in the county of Frederiksborg, Denmark.

1.4.3 Paper C

The focus in this paper is on modelling bivariate time series of counts. In the paper two models are considered. The basic structure of both models is a Poisson-based bivariate INGARCH model, which is capable of capturing the serial dependence between two time series of counts. The difference between the two models lies in the formulation of the bivariate conditional Poisson distribution. In the first model (originally proposed in Liu (2012, Chapter 4)) the Poisson distribution is constructed through trivariate reduction of independent Poisson random variables. In the second model, proposed in this paper, the Poisson distribution is constructed from two univariate Poisson distributions combined by use of a copula.

Mathematically, the models are formulated as follows. Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2})^T$ be the bivariate observation at time t where $\{Y_{t,1}, t \geq 1\}$ and $\{Y_{t,2}, t \geq 1\}$ are the two time series of interest. The common Poisson-based bivariate INGARCH model of order $(1,1)$ is defined as

$$\mathbf{Y}_t | \mathcal{F}_{t-1} \sim \mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi), \quad \boldsymbol{\lambda}_t = (\lambda_{t,1}, \lambda_{t,2})^T = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}, \tag{1.8}$$

where $\mathcal{F}_t = \sigma\{\boldsymbol{\lambda}_1, \mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$ is the σ-algebra of past events, $\varphi \in I_\varphi$ where $I_\varphi \subseteq \mathbb{R}$, $\boldsymbol{\delta} = (\delta_1, \delta_2) \in \mathbb{R}_+^2$ and $\mathbf{A} = \{\alpha_{ij}\}, \mathbf{B} = \{\beta_{ij}\}$ are both 2×2 matrices with non-negative entries. The notation $\mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi)$ represents a bivariate Poisson distribution whose marginal Poisson distributions have means $\lambda_{t,1}$ and $\lambda_{t,2}$, respectively, and φ is used for modelling the dependence between the two time series. The parameter vector $(\delta_1, \delta_2, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$ is denoted by $\boldsymbol{\theta}$.

For the first model the pmf of the bivariate Poisson distribution is given as

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}}(Y_{t,1} = m, Y_{t,2} = n | \mathcal{F}_{t-1}) \\ &\quad = e^{-(\lambda_{t,1} + \lambda_{t,2} - \varphi)} \frac{(\lambda_{t,1} - \varphi)^m}{m!} \frac{(\lambda_{t,2} - \varphi)^n}{n!} \sum_{s=0}^{\min\{m,n\}} \binom{m}{s} \binom{n}{s} s! \left(\frac{\varphi}{(\lambda_{t,1} - \varphi)(\lambda_{t,2} - \varphi)} \right)^s. \end{aligned} \tag{1.9}$$

For this model the conditional correlation between the two time series at a given time t is $\mathrm{Cor}(Y_{t,1}, Y_{t,2} | \mathcal{F}_{t-1}) = \varphi(\lambda_{t,1}\lambda_{t,2})^{-1/2}$.

The bivariate Poisson distribution of the second model is defined through its cdf, which is given as

$$\mathbb{P}_{\boldsymbol{\theta}}(Y_{t,1} \leq m, Y_{t,2} \leq n | \mathcal{F}_{t-1}) = C_\varphi(F_{\lambda_{t,1}}(m), F_{\lambda_{t,2}}(n)), \tag{1.10}$$

where C_φ is a copula parameterised by φ and F_λ is the cdf of a Poisson distribution with mean λ.

We notice that the conditional mean process $\{\boldsymbol{\lambda}_t\}$ of model (1.8) constitutes a Markov chain. Proposition C.2 (Liu, 2012, Proposition 4.2.1) gives conditions for $\{\boldsymbol{\lambda}_t\}$ to have a stationary distribution under the model (1.8) with pmf given by (1.9). In addition, the first part of the proposition provides conditions for the stationary distribution to be unique, and the second part of the proposition provides conditions for $\{\boldsymbol{\lambda}_t\}$ to be a geometric moment contracting Markov chain with a unique stationary and ergodic distribution. In section C.3 we remark that this result generalises to a setup where the bivariate Poisson distribution is replaced by a bivariate distribution constructed through trivariate reduction of variables with a distribution belonging to an exponential family satisfying a few regularity conditions. An example of a distribution that satisfies the conditions is the negative binomial distribution.

A natural way of estimating the parameters of model (1.8), with pmf given by (1.9), is MLE. Theorem C.4 states that the MLE of the parameter is strongly consistent. The result follows by adjusting the results of Wang et al. (2012) to this two-dimensional setup.

The topic of section C.4 is the model (1.8) with cdf given by (1.10). Proposition C.9 gives sufficient conditions on the copula in order for the result of Proposition C.2 to hold when the bivariate Poisson distribution is given by (1.10). These

conditions are satisfied for many Archimedean copulas, especially the Frank and the Clayton copula. The result of Lemma C.10 is that $\{\lambda_t\}$ still constitutes an e-chain (which is the case under (1.9) due to Liu (2012)) when the distribution is now given by (1.10). With this lemma the result of Proposition C.9 follows.

In the final two sections, the two types of models are compared through a simulation study and application to data. The results of these preliminary studies indicates that for a small sample size the copula-based model provides a better fit compared to the model proposed in Liu (2012).

The work presented in this paper should be considered as work in progress.

Bibliography

Aitchison, J. and C. Ho (1989). The multivariate Poisson-log normal distribution. *Biometrika 76*(4), 643–653.

Alzaga, J. M. S. and E. G. Déniz (2008). Construction of multivariate distributions: a review of some recent results. *SORT 32*(1), 3–36.

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics 5*(3), 151–157.

Beall, G. (1939). Methods of estimating the population of insects in a field. *Biometrika 30*(3/4), 422–439.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), 192–236.

Chatelain, F., S. Lambert-Lacroix, and J. Tourneret (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities. *Statistics and Computing 19*(3), 283–301.

Choo, L. and S. Walker (2008). A new approach to investigating spatial variations of disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(2), 395–405.

Clayton, D. and J. Kaldor (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics 43*(3), 671–681.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*(3), 729–737.

Cramér, H. and H. Wold (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society 1*(4), 290–294.

Darroch, J. N., S. L. Lauritzen, and T. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics 8*(3), 522–539.

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis 93*(1), 40–57.

Edwards, C. B. and J. Gurland (1961). A class of distributions applicable to accidents. *Journal of the American Statistical Association 56*(295), pp. 503–517.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association 81*(395), 709–721.

Fink, A. and M. Jodeit Jr (1984). On Chebyshev's other inequality. *IMS Lecture Notes-Monograph Series 5*, 115–120.

Fiocco, M., H. Putter, and J. Van Houwelingen (2009). A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics 10*(2), 245–257.

Genest, C. and J. Neslehova (2007). A primer on copulas for count data. *Astin Bulletin 37*(2), 475–515.

Grandell, J. (1997). *Mixed Poisson Process*, Volume 77 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Griffiths, R. (1984). Characterization of infinitely divisible multivariate gamma distributions. *Journal of Multivariate Analysis 15*(1), 13–20.

Griffiths, R. and R. K. Milne (1987). A class of infinitely divisible multivariate negative binomial distributions. *Journal of Multivariate Analysis 22*(1), 13–23.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson model. Technical Report 8113, MPRA.

Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance 14*(4), 564–583.

Henderson, R. and S. Shimakura (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika 90*(2), 355–366.

Jensen, J. L. (2011a). Asymptotic normality of m-estimators in nonhomogeneous hidden markov models. *Journal of Applied Probability 48A*, 295–306.

Jensen, J. L. (2011b). Central limit theorem for functions of weakly dependent variables. In *Proceeding of ISI-meeting in Dublin, 2011*. ISI.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis 94*(2), 401–419.

Johnson, N. L. and S. Kotz (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley New York.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. Wiley New York.

Kocherlakota, S. and K. Kocherlakota (1992). *Bivariate Discrete Distributions*, Volume 132 of *STATISTICS: textbooks and monographs*. New York: Marcel Dekker.

Krishnamoorthy, A. and M. Parthasarathy (1951). A multivariate gamma-type distribution. *The Annals of Mathematical Statistics 22*(4), 549–557.

Lauritzen, S. L. (2002). *Lectures on Contingency Tables*. Aalborg University.

Le Cessie, S. and J. C. Van Houwelingen (1994). Logistic regression for correlated binary data. *Applied Statistics 43*(1), 95–108.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics 80*(1), 221–39.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pp. i–163. JSTOR.

Liu, H. (2012). *Some Models for Time Series of Counts*. Ph. D. thesis, Columbia University.

Mardia, K. V. (1970). *Families of Bivariate Distributions*, Volume 27 of *Griffin's Statistical Monographs and Courses*. Griffin London.

McCullagh, P. and J. Møller (2006). The permanental process. *Advances in Applied Probability 38*(4), 873–888.

Møller, J. and E. Rubak (2010). A model for positively correlated count variables. *International Statistical Review 78*(1), 65–80.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer.

Nikoloulopoulos, A. K. and D. Karlis (2009). Modeling multivariate count data using copulas. *Communications in Statistics - Simulation and Computation 39*(1), 172–187.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis 110*, 4–18.

Rubak, E. (2010). *Likelihood Based Inference and Diagnostics for Spatial Data Models*. Ph. D. thesis, Aalborg University, The Faculty of Engineering and Science.

Rubak, E., J. Møller, and P. McCullagh (2010). Statistical inference for a class of multivariate negative binomial distributions. Technical Report R-2010-10, Aalborg University.

Shin, K. and R. Pasupathy (2007). A method for fast generation of bivariate Poisson random vectors. In *Simulation Conference, 2007 Winter*, pp. 472–479. IEEE.

Shin, K. and R. Pasupathy (2010). An algorithm for fast generation of bivariate Poisson random vectors. *INFORMS Journal on Computing 22*(1), 81–92.

Shirai, T. (2007). Remarks on the positivity of α-determinants. *Kyushu Journal of Mathematics 61*(1), 169–189.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris. 8*, 229–231.

Trivedi, P. K. and D. M. Zimmer (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics 1*(1), 1–111.

Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association 83*(401), 37–42.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis 92*(1), 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Vere-Jones, D. (1997). Alpha-permanents and their applications to multivariate gamma, negative binomial and ordinary binomial distributions. *New Zealand Journal of Mathematics 26*(1), 125–149.

Wang, C., H. Liu, J.-F. Yao, R. Davis, and W. Li (2012). Self-excited threshold Poisson autoregression. Submitted.

Paper

A

On the 3-dimensional negative binomial distribution

Camilla Mondrup Andreassen and Jens Ledet Jensen

Department of Mathematics, Aarhus University, Denmark

Abstract: We consider a particular generalisation of the negative binomial distribution to the multivariate case obtained through a specification of the probability generating function as the negative power of a certain polynomial. The probability function itself has previously been derived for the two-dimensional case only, and inference in the multivariate negative binomial distribution has hitherto been restricted to the use of composite likelihood based on one- or two-dimensional marginals. In this paper we derive the three-dimensional probability function as a sum with positive terms only and study the range of possible parameter values. We illustrate the use of the three-dimensional distribution for modelling three correlated SAR images.

Keywords: Correlated count data, inference, mixture model, multivariate analysis.

A.1 Introduction

Correlated count data are often encountered in applied fields. A simple example is when data are collected in a field, and where the 'living conditions' vary over the field. Counts at nearby positions are more similar than counts at distant positions due to this underlying variation. Another example, which we consider in detail in section A.3, is a series of SAR images taken before and after a major event. Here, the correlation of interest is between the images for each pixel of the images. In each pixel the number of photons received follows a mixed Poisson distribution, where the mixture variable represents the reflected intensity that varies due to the local structure in the area corresponding to the pixel. The above two examples point to the natural class of mixture models, where the counts are conditionally independent and Poisson distributed given the value of a latent variable. Correlation is then introduced through the latent variable. The mixture representation is a common motivation for the use of the univariate negative binomial distribution, the latter being a mixture with respect to a gamma distribution. The multivariate negative binomial model we consider in this paper can in some parts of the parameter space be obtained as a mixture distribution where the mixture distribution is multivariate gamma (Vere-Jones (1997), Henderson and Shimakura (2003) and Møller and Rubak (2010)). Other poisson mixture models are considered in Choo and Walker (2008) and in Aitchison and Ho (1989). In the latter the mixture distribution is a multivariate log-normal which allows for the extra flexibility of negative correlations as opposed to the case of a gamma mixture distribution.

The starting point for the class of multivariate negative binomial distributions of this paper is the probability generating function, given as a negative power of a polynomial of a certain form. The precise definition is given in (A.1) below. In particular, the form of the probability generating function implies that the one-dimensional marginals follows a negative binomial distribution. The class of multivariate negative binomial distributions has previously been studied in Griffiths (1984), Griffiths and Milne (1987), Vere-Jones (1997), Henderson and Shimakura (2003) and Møller and Rubak (2010). The probability function derived from the generating function can be expressed through the so-called α-permanents (Vere-Jones, 1997). No simple explicit form for the α-permanents is known and for this reason, inference in the multivariate negative binomial distribution has hitherto been restricted to the use of composite likelihood based on one- or two-dimensional marginals (Chatelain et al. (2009) and Rubak et al. (2010)). The two-dimensional probability function can be calculated as a sum where the number of terms equals the smaller of the two counts (the formula is given in (A.5) below). The two-dimensional probability function has been derived a number of times in the literature. One of the earliest references is Edwards and Gurland (1961). In that paper, as well as in Henderson and Shimakura (2003) a formula with an alternating sum is described. A sum formula with positive terms only, is given in Griffiths and Milne (1987) and Rubak et al. (2010). In this paper we extend these results and derive the three-dimensional probability function as a sum with positive terms only. The formula is obtained by a suitable expansion of the probability generating function and using the two-dimensional formula along the way. The resulting formula consists of a sum over four indices.

The multivariate negative binomial distribution is parameterised by a shape parameter $\alpha > 0$ and a matrix A. In the two-dimensional case A can be replaced by a symmetric matrix, and existence of the distribution at the parameter point (α, A) implies existence for all $(\tilde{\alpha}, A)$, $\tilde{\alpha} > 0$. In particular, this implies that in the two-dimensional case all distributions are infinite divisible. In the three-dimensional case these results do not hold, see Griffiths and Milne (1987) and Vere-Jones (1997). In this paper we study the subclass of infinite divisible distributions in order to obtain more explicit results. We also study subclasses of the infinite divisible distributions where simple descriptions of the parameter space can be given.

In summary, we focus in this paper on properties of the three-dimensional negative binomial distribution, and illustrate the use of this distribution for modelling three-dimensional count data. The paper is organised as follows. We start in section A.2 by defining the multivariate negative binomial model and stating and proving the results on existence in the two- and three-dimensional cases. The results for the two-dimensional case, as well as the existence result for the three-dimensional case, are known, but we give here simple and self-contained proofs, and use the proofs for the known results for deriving the new results. In section A.3 we apply likelihood estimation based on the three-dimensional probability function to a real data example.

A.2 The multivariate negative binomial distribution

In this section we state and prove results for the multivariate negative binomial distribution, which we then use in the data analysis in section A.3. Some of these results are known, but we give here simple and self-contained proofs.

Let $N = (N_1, \ldots, N_m)$ be an m-dimensional positive discrete random vector, with probability generating function (pgf) of the form

$$\varphi(z) = \mathbb{E} \prod_{i=1}^{m} z_i^{N_i} = |I + (I - Z)A|^{-\alpha}, \tag{A.1}$$

with $|\cdot|$ the determinant, $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$, $Z = \operatorname{diag}(z_1, \ldots, z_m)$, I the $m \times m$ identity matrix, $\alpha > 0$ and $A = \{a_{ij}\}$ a real $m \times m$ matrix. The possible values of the matrix A is the subject of sections A.2.1 and A.2.2 below and the interpretation of A in terms of means and covariances is given below. The generating function of N_i is obtained on setting $z_j = 1$, $j \neq i$, that is, $\{1 + (1 - z_i)a_{ii}\}^{-\alpha}$. This is the generating function of a negative binomial distribution with mean αa_{ii}. Due to this fact we refer to the distribution corresponding to (A.1) as the multivariate negative binomial (MNB) distribution and denote the distribution by $\mathrm{NB}_m(\alpha, A)$. For the density we use the notation $p_m(n_1, \ldots, n_m) = \mathbb{P}(N_1 = n_1, \ldots, N_m = n_m)$. The model has been studied in detail in Griffiths and Milne (1987) and Vere-Jones (1997). Other models of this type have been studied in Griffiths (1984), Henderson and Shimakura (2003), Chatelain et al. (2009), Møller and Rubak (2010) and Rubak et al. (2010).

When using the MNB distribution for data analysis the parameterisation by A is of interest since the moments can be found as simple functions of α and A. In

particular we have

$$\begin{aligned} \mathbb{E}\, N_i &= \alpha a_{ii}, & \mathbb{E}(N_i N_j) &= \alpha^2 a_{ii} a_{jj} + \alpha a_{ij} a_{ji}, \quad i \neq j, \\ \operatorname{Var} N_i &= \alpha a_{ii}^2 + \alpha a_{ii}, & \operatorname{Cov}(N_i, N_j) &= \alpha a_{ij} a_{ji} \quad i \neq j. \end{aligned} \tag{A.2}$$

The correlation can then be found, and using Proposition A.1 below we obtain an upper limit for the correlation which, in the case $a_{ii} \leq a_{jj}$, becomes

$$\operatorname{Cor}(N_i, N_j) = \frac{a_{ij} a_{ji}}{\sqrt{a_{ii}(a_{ii}+1) a_{jj}(a_{jj}+1)}} \leq \sqrt{\frac{a_{ii}(a_{jj}+1)}{a_{jj}(a_{ii}+1)}} = \sqrt{\frac{(\mathbb{E}\, N_i)^2 \operatorname{Var} N_j}{(\mathbb{E}\, N_j)^2 \operatorname{Var} N_i}}. \tag{A.3}$$

The marginal generating function for a subset of variables is obtained from (A.1) on setting $z_j = 1$ for those j's not in the subset. It follows that the marginal generating function has the same form as in (A.1), with the same α and with A replaced by the submatrix obtained by deleting rows and columns corresponding to those variables not considered.

Whereas the representation in (A.1) is well suited for moment properties it is less suited when one wants to find the probability function and find necessary and sufficient conditions on the matrix A for the model to exist. Instead we use the reparameterisation obtained by setting $Q = A(I+A)^{-1} = I - (I+A)^{-1}$ and rewriting (A.1) as

$$\varphi(z) = \left(\frac{|I - Q|}{|I - ZQ|} \right)^{\alpha}. \tag{A.4}$$

In order that $\varphi(z)$ from (A.1) represents a pgf we must have $|I + (I - Z)A| > 0$ for $Z = \operatorname{diag}(z_1, z_2, \ldots, z_m)$, $0 \leq z_i \leq 1$, which shows that $|I + A| > 0$, allowing us to define Q. Similarly, for (A.4) to be a pgf we must have $|I - Q| > 0$, and A is given through the relation $A = Q(I - Q)^{-1}$.

To study when (A.4) is a pgf of a positive discrete random vector we have to consider for which matrices A the function (A.4) is defined for $|z_i| \leq 1$, $i = 1, \ldots, m$, and when all the coefficients of a power series expansion are non-negative. We study this question in the two- and three-dimensional cases in the following subsections. In the general case Vere-Jones (1997) has expressed the probability function $p_m(n_1, \ldots, n_m)$ through the α-permanent of a $n \times n$ matrix with $n = n_1 + \cdots + n_m$. The α-permanent itself is expressed as a sum over all permutations of n elements, where each term in the sum involves the number of cycles in the permutation. This formula is not useful for numerical implementation, and for this reason we concentrate in this paper on the two- and three-dimensional distributions.

A.2.1 Two dimensions

In this subsection we deduce necessary and sufficient conditions for (A.1) and (A.4) to be well defined in the two-dimensional case and we find an expression for the probability function by use of a Taylor series. Furthermore, we consider for which matrices Q the distribution corresponding to (A.4) is infinitely divisible. We use the notation $\alpha^{\uparrow k} = \alpha(\alpha+1)\cdots(\alpha+k-1) = \Gamma(\alpha+k)/\Gamma(\alpha)$. The results are as follows.

Proposition A.1. *Let $\varphi(z) = |I - Q|^{\alpha}|I - ZQ|^{-\alpha}$ with $\alpha > 0$ and $Q = \{q_{ij}\}$ a real 2×2 matrix. Then φ is a pgf for a random variable in $\mathbb{N}^2$ if and only if*

$$0 \leq q_{ii} < 1,\, i = 1, 2, \quad \textit{and} \quad 0 \leq q_{12}q_{21} < (1 - q_{11})(1 - q_{22}).$$

The probability function can be written as

$$p_2(r, s) = |I - Q|^{\alpha} \frac{\alpha^{\uparrow r}\alpha^{\uparrow s}}{r!s!} \sum_{k=0}^{\min\{r,s\}} \binom{r}{k}\binom{s}{k} \frac{k!}{\alpha^{\uparrow k}} (q_{12}q_{21})^k q_{11}^{r-k} q_{22}^{s-k}, \tag{A.5}$$

for $r, s \in \mathbb{N}$. The equivalent conditions in term of $A = Q(I - Q)^{-1}$ are $a_{ii} \geq 0$, $i = 1, 2$, and $0 \leq a_{12}a_{21} \leq a_{11}a_{22} + \min\{a_{11}, a_{22}\}$.

Proof. Let φ be a pgf Then

$$\varphi(z) = \sum_{r,s \geq 0} p_2(r, s) z_1^r z_2^s, \tag{A.6}$$

and $p_2(0, 0) = \varphi(0, 0) = |I - Q|^{\alpha} \geq 0$, but the form of φ implies that, actually, $|I - Q|^{\alpha} > 0$. Let

$$\eta(z_1, z_2) = |I - ZQ| = (1 - q_{11}z_1)(1 - q_{22}z_2) - q_{12}q_{21}z_1z_2.$$

In order for $\varphi(z)$ to be finite for $|z_1|, |z_2| \leq 1$, we see that $\eta(z_1, z_2) > 0$ in this region. In particular, $|I - Q| = \eta(1, 1) > 0$ implies that $q_{12}q_{21} < (1 - q_{11})(1 - q_{22})$. Next, $\eta(z_1, 0) = 1 - q_{11}z_1 > 0$ for $|z_1| \leq 1$ implies that $q_{11} < 1$. Also, since $\varphi(z_1, 0) = \sum_{r=0}^{\infty} p_2(r, 0) z_1^r$ is non-decreasing for $z_1 > 0$, we find that $q_{11} \geq 0$. By symmetry we get $0 \leq q_{22} < 1$.

Finally, assume that $q_{12}q_{21} < 0$. Let $a, b > 0$ and let $\varphi(az, bz) < \infty$ for $0 < z < z_0$. Then $\varphi(az, bz)$ is non-decreasing for $0 < z < z_0$ and therefore $\eta(az, bz)$ has to be non-increasing. When $q_{12}q_{21} < 0$ it follows that $\eta(az, bz)$ is increasing for sufficiently large z, and therefore there must exist $z_0 < \infty$ with $\eta(az_0, bz_0) = 0$. If $q_{11}q_{22} > 0$ we take $a = 1/q_{11}$ and $b = 1/q_{22}$, and then

$$\eta(az, bz) = (1 - z)^2 - q_{12}q_{21}(q_{11}q_{22})^{-1}z^2 > 0,$$

for all $z > 0$. Thus, this contradicts the assumption $q_{12}q_{21} < 0$. Similarly, if $q_{11} = 0$, $q_{22} > 0$ we take $a = -q_{22}(q_{12}q_{21})^{-1}$ and $b = 1/q_{22}$, with similar choices in the case $q_{11} > 0$, $q_{22} = 0$, and if $q_{11} = q_{22} = 0$ we take $a = b = 1$. In all cases we get a contradiction, which proves that $q_{12}q_{21} \geq 0$.

For the sufficiency part we see that when $0 \leq q_{12}q_{21} < (1 - q_{11})(1 - q_{22})$ we can expand $\eta(z_1, z_2)^{-\alpha}$ in powers of $q_{12}q_{21}z_1z_2$. Thus

$$\eta(z_1, z_2)^{-\alpha} = \sum_{k=0}^{\infty} \frac{\alpha^{\uparrow k}}{k!} (z_1z_2q_{12}q_{21})^k (1 - z_1q_{11})^{-(\alpha+k)} (1 - z_2q_{22})^{-(\alpha+k)}.$$

When $0 \leq q_{ii} < 1$ we can expand $(1 - q_{ii}z_i)^{-(\alpha+k)}$ in powers of $q_{ii}z_i$, $i = 1, 2$. We then obtain

$$\eta(z_1, z_2)^{-\alpha} = \sum_{k=0}^{\infty} \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \frac{\alpha^{\uparrow k}}{k!} \frac{(\alpha + k)^{\uparrow m_1}}{m_1!} \frac{(\alpha + k)^{\uparrow m_2}}{m_2!} (q_{12}q_{21})^k q_{11}^{m_1} q_{22}^{m_2} z_1^{k+m_1} z_2^{k+m_2},$$

which shows that the coefficient to $z_1^r z_2^s$ in (A.6) is given by (A.5). Since the coefficients are all non-negative it follows that φ is a pgf. □

One of the first appearances of the two-dimensional probability function is in Edwards and Gurland (1961). In that paper, as well as in Henderson and Shimakura (2003), a formula with an alternating sum is described. A formula similar to (A.5) can be found in Griffiths and Milne (1987) and in Rubak et al. (2010). The conditions for existence in the above proposition correspond to the ones in Proposition 5.6 in Vere-Jones (1997). Our proof is based on the characteristics of a pgf and is different from the one given by Vere-Jones. The reason for including the proof in this paper is that the proof is used directly when the three-dimensional probability function is derived in the next subsection.

From (A.5) it is seen that the distribution depends on q_{12} and q_{21} through $q_{12}q_{21}$ only (equivalently the distribution depends on a_{12} and a_{21} through $a_{12}a_{21}$ only). Thus Q, or A, can always be chosen as a symmetric matrix (this is not true for the NB_3-distribution). Furthermore, the condition $a_{12}a_{21} \geq 0$ shows that the correlation between the two variables is always positive. Hence, the model can only be used to analyse data where a positive correlation is expected. This is the case for the data we analyse in section A.3.

By definition the $\mathrm{NB}_2(\alpha, A)$-distribution is infinite divisible if and only if (A.1), with α replaced by α/n, is a pgf for all $n \geq 1$. Since the conditions in Proposition A.1 do not involve α, we see that indeed all $\mathrm{NB}_2(\alpha, A)$-distributions are infinite divisible. For general m we have that the $\mathrm{NB}_m(\alpha, A)$-distribution is an infinite divisible distribution if and only if the $\mathrm{NB}_m(\tilde{\alpha}, A)$-distribution exists for all $\tilde{\alpha} > 0$ (see the argument leading to Proposition 3.11 in Vere-Jones (1997)).

A.2.2 Three dimensions

In this subsection we consider the case $m = 3$. We give necessary conditions for (A.1) and (A.4) to be well defined and we find an expression for the probability function by use of a Taylor series. We deduce necessary and sufficient conditions on Q for the distribution corresponding to (A.4) to exist for all $\alpha > 0$. In special cases we express the corresponding conditions on A.

Contrary to the two-dimensional case not all $\mathrm{NB}_3(\alpha, A)$-distributions are infinite divisible, that is, existence of the $\mathrm{NB}_3(\alpha, A)$-distribution does not imply existence of the $\mathrm{NB}_3(\tilde{\alpha}, A)$-distribution for all $\tilde{\alpha} > 0$. For a fixed α there does not seem to exist simple necessary and sufficient conditions on A for existence of $\mathrm{NB}_3(\alpha, A)$. However, requiring existence for all $\alpha > 0$ there is a simple characterisation. The necessary and sufficient conditions in the proposition below have earlier been given in Theorem 2 of Griffiths and Milne (1987) and in Proposition 5.7 of Vere-Jones (1997). The proposition is formulated in terms of Q, but as in the two-dimensional case it can be reformulated in terms of A if this is needed.

Proposition A.2. *Let $\varphi(z) = |I - Q|^{\alpha}|I - ZQ|^{-\alpha}$ with $Q = \{q_{ij}\}$ a real 3×3 matrix. Then φ is a pgf for all $\alpha > 0$ for a random variable in $\mathbb{N}^3$ if and only if*

1. $|I - Q| > 0$,
2. $0 \leq q_{ii} < 1$, $i = 1, 2, 3$,
3. $0 \leq q_{ij}q_{ji} < (1 - q_{ii})(1 - q_{jj})$, $i \neq j$,
4. $q_{12}q_{23}q_{31}, q_{13}q_{32}q_{21} \geq 0$.

The probability function can be written as

$$p_3(s,t,u) = |I-Q|^{\alpha}\frac{\alpha^{\uparrow s}\alpha^{\uparrow t}\alpha^{\uparrow u}}{s!t!u!} \times \sum_{k=0}^{\min\{s,t\}}\binom{s}{k}\binom{t}{k}\frac{k!}{\alpha^{\uparrow k}}\sum_{m=0}^{\min\{s+t,u\}}\binom{u}{m}\frac{m!}{\alpha^{\uparrow m}}q_{33}^{u-m}h(s-k,t-k,m,k), \tag{A.7}$$

for $s,t,u \in \mathbb{N}$, where $h(\cdot)$ is given below through (A.8), (A.9) *and* (A.10) *with $b_{ij} = q_{ij}q_{ji}$ and $b_{ijk} = q_{ij}q_{jk}q_{ki}$.*

Proof. Let

$$\begin{aligned}\eta(z_1,z_2,z_3) &= |I-ZQ| \\ &= (1-z_1q_{11})(1-z_2q_{22})(1-z_2q_{33}) - z_1z_2z_3(q_{12}q_{23}q_{31}+q_{13}q_{32}q_{21}) \\ &\quad - z_1z_2q_{12}q_{21}(1-z_3q_{33}) - z_1z_3q_{13}q_{31}(1-z_2q_2) - z_2z_3q_{23}q_{32}(1-z_1q_{11}).\end{aligned}$$

As in the proof for the two-dimensional case we obtain the conditions $|I-Q|>0$, $0 \le q_{ii} < 1$ and $0 \le q_{ij}q_{ji} < (1-q_{ii})(1-q_{jj})$, $i \ne j$, by considering $\eta(1,1,1)$, $\eta(z_1,0,0)$, $\eta(z_1,1,0)$, $\eta(az,bz,0)$ etc.

The necessity of $b_{ijk} = q_{ij}q_{jk}q_{ki} \ge 0$ follows from the condition $p_3(1,1,1) \ge 0$ for all $\alpha > 0$. The latter is equivalent to $\alpha^{-1}\frac{\partial^3}{\partial z_1\partial z_2\partial z_3}\eta(z)^{-\alpha}\big|_{z=0} \ge 0$. Letting α tend to zero we get $b_{123}+b_{132} \ge 0$. From $b_{ij} = q_{ij}q_{ji} \ge 0$, $i \ne j$, one sees that $b_{123}b_{132} \ge 0$ so that we obtain $b_{123}, b_{132} \ge 0$. Hence, the necessity has been proved.

For the sufficiency part we find by direct calculations that

$$\eta(z_1,z_2,z_3) = (1-q_{33}z_3)\{(1-H_1z_1)(1-H_2z_2) - H_{12}z_1z_2\},$$

where

$$H_1 = q_{11} + \frac{q_{13}q_{31}z_3}{1-q_{33}z_3}, \qquad H_2 = q_{22} + \frac{q_{23}q_{32}z_3}{1-q_{33}z_3},$$
$$H_{12} = q_{12}q_{21} + \frac{(q_{12}q_{23}q_{31}+q_{13}q_{32}q_{21})z^3}{1-q_{33}z_3} + \frac{q_{13}q_{31}q_{23}q_{32}z_3^2}{(1-q_{33}z^3)^2}.$$

Clearly, $H_1, H_2, H_{12} \ge 0$ for $0 \le z_3 \le 1$. Also, from $\eta(1,0,z_3) > 0$, $\eta(0,1,z_3) > 0$ and $\eta(1,1,z_3) > 0$ it follows that $H_1 < 1$, $H_2 < 1$ and $H_{12} < (1-H_1)(1-H_2)$. Hence, we can use the result from the two-dimensional case and we obtain that

$$\eta(z_1,z_2,z_3)^{-\alpha} = (1-q_{33}z_3)^{-\alpha}\sum_{r,s=0}^{\infty}\frac{\alpha^{\uparrow r}\alpha^{\uparrow s}}{r!s!}\sum_{k=0}^{\min(r,s)}\binom{r}{k}\binom{s}{k}\frac{k!}{\alpha^{\uparrow k}}H_1^{r-k}H_2^{s-k}H_{12}^k z_1^r z_2^s.$$

By using the binomial formula to expand H_1^{r-k} and H_2^{s-k} we obtain

$$H_1^{r-k}H_2^{s-k} = \sum_{m=0}^{r+s-2k}\left(\frac{z_3}{1-q_{33}z_3}\right)^m f(r-k,s-k,m),$$

with

$$f(a,b,m) = \sum_{v=\max(0,m-b)}^{\min(m,a)}\binom{a}{v}\binom{b}{m-v}b_{13}^v b_{23}^{m-v}q_{11}^{a-v}q_{22}^{b-m+v}. \tag{A.8}$$

Similarly, using the multinomial formula we obtain

$$H_{12}^k = \sum_{m=0}^{2k} \left(\frac{z_3}{1 - q_{33} z_3} \right)^m g(k, m),$$

with

$$g(k,m) = \sum_{v=\max(0,m-k)}^{\lfloor m/2 \rfloor} \binom{k}{v, m-2v, k-m+v} (b_{13} b_{23})^v (b_{123} + b_{132})^{m-2v} b_{12}^{k-m+v}, \quad \text{(A.9)}$$

with $\lfloor m/2 \rfloor$ denoting the integer part of $m/2$. Combining these two expressions we obtain

$$\eta(z)^{-\alpha} = (1 - q_{33} z_3)^{-\alpha} \sum_{r,s=0}^{\infty} \left\{ \frac{\alpha^{\uparrow r} \alpha^{\uparrow s}}{r! s!} z_1^r z_2^s \right.$$
$$\left. \times \sum_{k=0}^{\min(r,s)} \binom{r}{k} \binom{s}{k} \frac{k!}{\alpha^{\uparrow k}} \sum_{m=0}^{r+s} \left(\frac{z_3}{1 - q_{33} z_3} \right)^m h(r-k, s-k, m, k) \right\},$$

with

$$h(a,b,m,k) = \sum_{c=\max(0,m-a-b)}^{\min(m,2k)} f(a,b,m-c) g(k,c). \quad \text{(A.10)}$$

Finally, we insert the expansion

$$(1 - q_{33} z_3)^{-\alpha} \left(\frac{z_3}{1 - q_{33} z_3} \right)^m = \sum_{u=0}^{\infty} \frac{(m+\alpha)^{\uparrow u}}{u!} q_{33}^u z_3^{m+u},$$

and finding the coeficient to $z_1^r z_2^s z_3^t$ in $\eta(z)^{-\alpha}$ we obtain (A.7). □

When deriving the probability function for the NB_3-model we used the representation of the two-dimensional density. This points to the possibility that generally higher order probability functions can be found from lower orders. However, this is likely not of much practical interest as the number of terms in the sum representing the probability function will grow very fast with the counts.

To illustrate the three-dimensional probability function we have created plots of (A.7) by calculating the probability as a function of (N_1, N_3) for different values of N_2 and Q. We consider two parameter matrices

$$A_1 = \begin{pmatrix} d_1 & a & e \\ a & d_2 & a \\ e & a & d_1 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} d_1 & \frac{1}{\rho_1} a & e \\ \rho_1 a & d_2 & \rho_2 a \\ e & \frac{1}{\rho_2} a & d_1 \end{pmatrix}, \quad \text{(A.11)}$$

that give rise to the same correlation matrix and the same marginal means. For the plots we have used the values $d_1 = 2$, $d_2 = 1$, $a = 1$, $e = 1.5$, $\rho_1 = 2$, $\rho_2 = 3.9$ and $\alpha = 1$. The plots can be seen in Figure A.1. In the three topmost plots the probabilities are based on A_1 and in the three lowermost plots the probabilities

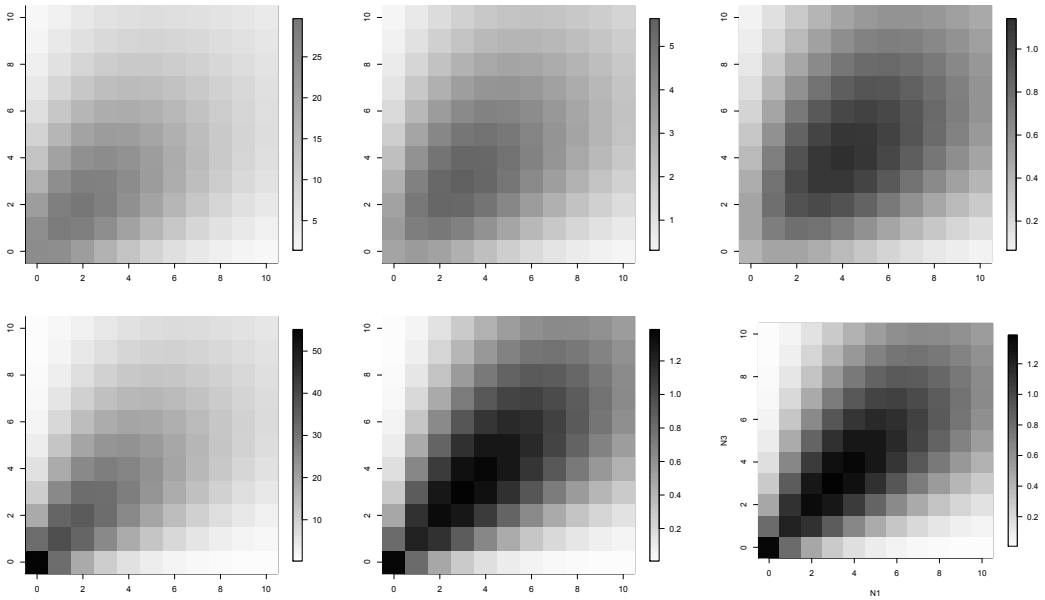


Figure A.1: Illustration of the three-dimensional probability function. The three topmost figures are based on A_1 and the three lowermost plots are based on A_2 in (A.11). The parameters are $d_1 = 2$, $d_2 = 1$, $a = 1$, $e = 1.5$, $\rho_1 = 2$, $\rho_2 = 3.9$ and $\alpha = 1$ and all probabilities have been multiplied by 10^6. From left to right the value of N_2 is 8, 10, and 12. For each column the same colour scale has been used.

are based on A_2. Even though there is the same mean and correlation structure in the two distributions, there is a clear difference in the two rows and hence in the distributions.

As mentioned above, no necessary and sufficient conditions for existence seem to be known, for a fixed α. The necessary conditions $a_{ij}a_{ji} \leq a_{ii}a_{jj} + \min\{a_{ii}, a_{jj}\}$, $|I + A| > 0$ and $a_{ii} \geq 0$ follow directly from Proposition A.1. Sufficient conditions for existence (for a general m) have been given in Griffiths and Milne (1987), Vere-Jones (1997) and Shirai (2007).

In Proposition A.3 below we consider conditions on A for (A.1) to represent an infinitely divisible distribution when A is symmetric. We notice that not all distributions in the NB_3 class can be represented by symmetric matrices A. Since A is symmetric if and only if Q is symmetric we can argue in terms of Q. For a symmetric Q we have $q_{12}q_{23}q_{31} + q_{13}q_{32}q_{21} = 2\sqrt{q_{12}q_{21}}\sqrt{q_{13}q_{31}}\sqrt{q_{23}q_{32}}$. Considering

$$Q = \begin{pmatrix} q & \beta & 0 \\ 0 & q & \beta \\ \beta & 0 & q \end{pmatrix}, \quad 0 \leq q < 1, 0 \leq \beta < 1 - q,$$

we get a distribution that satisfies the conditions in Proposition A.2, but not satisfying the equation above. This is contrary to the two-dimensional case where all the distributions can be obtained from symmetric matrices. However, the correlation matrix, see (A.3), can always be obtained from a symmetric matrix. Therefore,

choosing A to be symmetric will only have an influence on the moments of order three or higher and this is seldom of interest from a practical point of view.

Proposition A.3. *Let A be a symmetric matrix of the form*

$$A = \begin{pmatrix} d_1 & a_{12} & a_{13} \\ a_{12} & d_2 & a_{23} \\ a_{13} & a_{23} & d_3 \end{pmatrix},$$

and define for $i = 1, 3$, with $a_{32} = a_{23}$, $B_i = (1 + d_i)(1 + d_2) - a_{i2}^2$, $C_i = d_i(1 + d_2) - a_{i2}^2$, $D_i = d_2(1+d_i) - a_{i2}^2$ and $h = a_{13} - H$ with $H = a_{12}a_{23}/(1+d_2)$. Then $\varphi(z) = |I-(I-Z)A|^{-\alpha}$, $\alpha > 0$, is a pgf of a positive, discrete, infinitely divisible distribution if and only if

(i) $d_i \geq 0$, $i = 1, 2, 3$, $a_{ij}^2 \leq d_i d_j + \min\{d_i, d_j\}$, $i < j$,

(ii) $h^2(1 + d_2)^2 \leq \min\{B_3C_1, B_1C_3\}$, $(hd_2 - H)^2 \leq D_1D_3$,

(iii) $h = 0$ *or* $a_{12} = a_{23} = 0$ *or both* $a_{12}a_{23} \neq 0$ *and* $0 \leq \frac{h}{H} \leq \min\{\frac{B_1}{a_{12}^2}, \frac{B_3}{a_{23}^2}\}$.

Proof. The conditions in (i) follow from the two-dimensional case in Proposition A.1. Writing $Q = I - (I + A)^{-1}$ in terms of d_i and a_{ij} we use Proposition A.2. The condition $q_{ii} < 1$ is satisfied under assumption (i), and $q_{ii} \geq 0$ leads to (ii). Also (i) and (ii) gives $|I - Q| > 0$. The condition $q_{ij}q_{ji} < (1 - q_{ii})(1 - q_{jj})$ is in the present setting a consequence of the condition $q_{ii} \geq 0$. Finally, the condition $q_{12}q_{23}q_{31} \geq 0$ reduces to $h\{a_{12}B_3(1 + d_2)^{-1} - ha_{23}\}\{a_{23}B_1(1 + d_2)^{-1} - ha_{12}\} \geq 0$, which gives (iii). □

For later use in section A.3, we state the following corollary.

Corollary A.4. *Let the assumptions be as in Proposition A.3, but with $a_{12} = a_{23}$, $d_1 = d_3$. Then $\varphi(z) = |I + (I - Z)A|^{-\alpha}$ is a pgf of a positive discrete infinitely divisible distribution if and only if either $d_2 > 0$, $a_{12}^2 < d_1d_2 + \min\{d_1, d_2\}$ and*

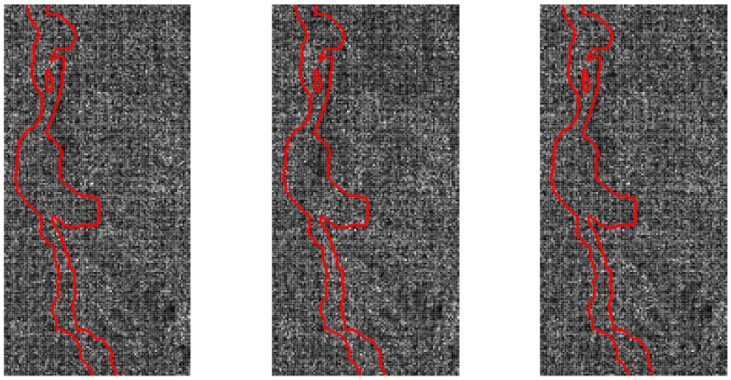
$$\max\Big\{\tfrac{a_{12}^2}{1+d_2}, \tfrac{2a_{12}^2-(1+d_1)d_2}{d_2}\Big\} \leq a_{13} \leq \tfrac{a_{12}^2+\sqrt{(d_1(1+d_2)-a_{12}^2)((1+d_1)(1+d_2)-a_{12}^2)}}{1+d_2}.$$

or both $d_2 = a_{12} = 0$ and $a_{13}^2 \leq d_1(1 + d_1)$.

Proof. This follows directly from Proposition A.3. □

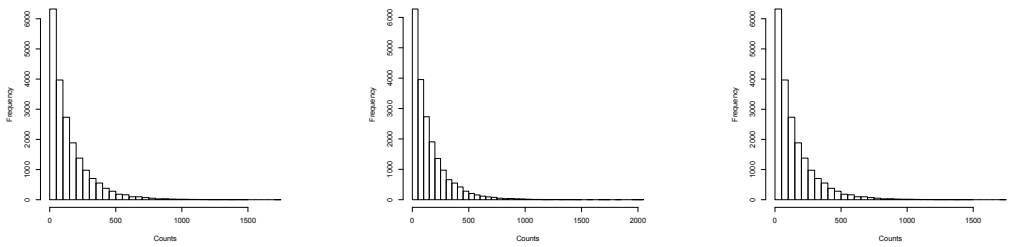
A.3 Application to data

To illustrate the use of the three-dimensional distribution, we consider the data from Chatelain et al. (2009). The data consists of three low-flux synthetic aperture radar (SAR) images: a reference image of the Nyiragongo volcano in Congo before an eruption and two secondary images of the same scene acquired after the eruption. Each of the images can be represented as a matrix with each entrance in the matrix corresponding to the measured number of photons in the corresponding pixel of the image. The data also contain a binary image (the mask), referred to as the ground truth in Chatelain et al. (2009), indicating the pixels of the image, which



(a) After (N_1) (b) Before (N_2) (c) After (N_3)

Figure A.2: Low-flux 200×100 radarsat images of the Nyiragongo volcano before and after an eruption. The pixels within the strip bounded by the red lines are the pixels that have been affected by the eruption.



(a) After (N_1) (b) Before (N_2) (c) After (N_3)

Figure A.3: Histograms of the observed counts for each image.

have been affected by the eruption. The low-flux images can be seen in Figure A.2. Histograms of the observed counts for each image can be seen in Figure A.3.

For each pixel three numbers of photons are observed corresponding to the three images. These will be denoted (N_1, N_2, N_3) where N_2 corresponds to the reference image and N_1 and N_3 correspond to the two secondary images. The individual pixel count originates from the reflectance of the radar signal by the ground, and the intensity of the reflectance is the mean of the poisson count for that pixel. The reflected intensity is random due to the micro structure of the area covered by the pixel. For a given type of ground the reflected intensities for different pixels are independent, and so the pixel counts within an image becomes independent. However, for the same pixel across a series of images, we expect a positive correlation since the pixel corresponds to the same area across the series of images. As a model for the three-dimensional data for each pixel we use the NB_3-distribution. When the type of ground changes the parameters of the distribution changes as well. We assume that the ground is locally homogeneous so that we can use the same distribution within a small window. Using the NB_3-distribution we can from the probability function in Proposition A.2 establish the likelihood

function and consider maximum likelihood estimation (MLE). Chatelain et al. (2009) analyse the data by using the NB_2-model and a composite likelihood. They furthermore consider estimation based on the method of moments, and perform a simulation study that shows that their proposed estimator based on the NB_2-distribution outperforms the moment estimator.

When specifying the model we consider the correlation structure in the data. Since image 1 and 3 both have been registered after the eruption, it is natural to assume that there is a stronger correlation between these two images than between image 1 and 2 and image 2 and 3. Hence, we assume that $r_{12} = r_{23} \leq r_{13}$ with $r_{ij} = \mathrm{Cor}(N_i, N_j)$. We let $r = r_{12} = r_{23}$. Furthermore, it is natural to assume that $\mathbb{E}N_1 = \mathbb{E}N_3$ since both images are obtained after the eruption. Since the correlation between image i and j, r_{ij}, is given by (A.3), which depends on a_{ij} and a_{ji} through $a_{ij}a_{ji}$ only, we restrict attention to symmetric matrices A with non-negative off-diagonal entries. Hence, we make the assumption that

$$A = \begin{pmatrix} a_{11} & \sqrt{rA_{12}} & \sqrt{r_{13}a_{11}(a_{11}+1)} \\ \sqrt{rA_{12}} & a_{22} & \sqrt{rA_{12}} \\ \sqrt{r_{13}a_{11}(a_{11}+1)} & \sqrt{rA_{12}} & a_{11} \end{pmatrix}, \tag{A.12}$$

where $A_{12} = \sqrt{a_{11}(a_{11}+1)a_{22}(a_{22}+1)}$. The matrix A is then of the type considered in Corollary A.4.

Moreover, it is assumed that $\alpha = 1$. This is due to the experimental setup as explained in Chatelain et al. (2009). A check of this assumption can be based on the relation $\mathrm{Var}\, N_i = \mathbb{E}N_i + \alpha^{-1}(\mathbb{E}N_i)^2$. Using non-overlapping windows with n pixels, indexed by i, we calculate the average $\bar{N}_i$ and the empirical variance s_i^2 and use the unbiased estimating equation

$$\sum_i \left\{ s_i^2 - \bar{N}_i - \tfrac{1}{1+n\alpha}(n\bar{N}_i^2 - \bar{N}_i) \right\} = 0.$$

For non-overlapping windows of size 2×2 the estimate is $\hat{\alpha} = 0.951$ with asymptotic standard error 0.028. Based on this analysis it seems reasonable to assume that $\alpha = 1$.

For each pixel we use an $n \times n$ window centered at the pixel and estimate the correlation r. If the estimated value $\hat{r}$ is below some threshold t the pixel is classified as having a change from before to after the eruption. Since pixels are considered independent we have n^2 observations from the NB_3-distribution to be used for the estimation.

We use the NB_3-model with A given by (A.12) and require infinite divisibility, such that the parameter space is given by Corollary A.4. To reduce the computational cost we estimate the means a_{11}, a_{22} and a_{33} from the marginal distributions (recall that $\alpha = 1$). Thus, we use MLE for estimating r and r_{13}. We consider estimation windows of size 3×3, 5×5 and 7×7.

In order to estimate the parameters r and r_{13} by use of MLE we have to calculate the probability given in (A.7) for every pixel for several values of the parameters. Even though the formula by itself is simple, the computational complexity increases rapidly with the size of the counts. Here, complexity refers to the number of terms that has to be calculated. We notice that formula (A.7) has the lowest

number of terms if $s \leq u \leq t$. Figure A.3 shows histograms of the observed counts for each image and it is seen that the counts lie in the interval $[0, 2017]$. If we consider the vector of averages $(\tilde{N}_{(1)}, \tilde{N}_{(3)}, \tilde{N}_{(2)}) \approx (147, 151, 148)$ formula (A.7) has 1 813 444 terms that needs to be calculated, and if we consider the vector of medians $(\tilde{N}_{(1)}, \tilde{N}_{(3)}, \tilde{N}_{(2)}) = (96, 97, 96)$ there are 761 838 terms. In the lower end, if we consider the vector $(5, 23, 6)$, corresponding to one of the pixels, there are only 1434 terms. In order to be able to perform the calculations we use a saddlepoint approximation (Jensen, 1995) to (A.7) whenever the number of terms to be calculated is too large. We notice that it follows from (A.7) that the number of terms is bounded by $2(N_{(1)} + 1)^2(N_{(2)} + 1)$ and that this bound is independent of $N_{(3)}$. Therefore we use the saddlepoint approximation whenever at least two of the values of N_1, N_2 and N_3 are greater than 10 giving the upper bound 29 282 for the number of terms when calculating (A.7). Hence, for 930 pixels we use the exact probability given in (A.7), for 16 028 pixels we use a full saddlepoint approximation to (A.7) and for the rest we use a saddlepoint approximation to $p_3(n_1, \cdot, \cdot)$, $p_3(\cdot, n_2, \cdot)$ and $p_3(\cdot, \cdot, n_3)$ respectively. We defer the details to the appendix.

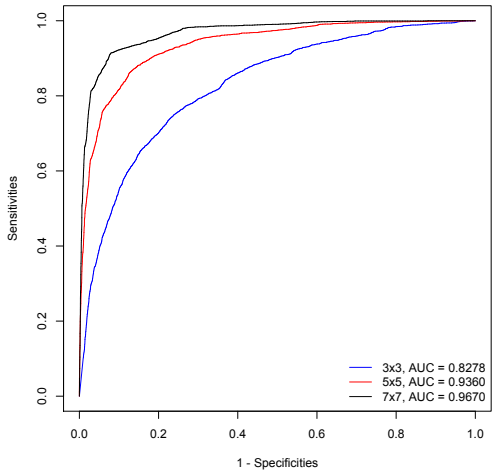
As a measure of the quality of the detection algorithm we use the area under the receiver operating characteristic (ROC) curve (AUC) (Fawcett, 2006). The ROC curve is a way of illustrating the perfomance of a binary classifier. It is created by plotting the fraction of true positives out of the positives versus the fraction of false positives out of the negatives, at various threshold settings. The AUC can therefore be used as a measure of the quality of the classifier. The result is seen in Figure A.4(a). We find that the AUC increases with increasing window size. As an alternative to the ROC curves we consider the mean and standard deviation (std) of $\hat{r}$ and $\hat{r}_{13}$ when the estimation is based on a 3×3 window. For the affected pixels the mean of $\hat{r}$ is 0.337 with a std of 0.240 and the mean of $\hat{r}_{13}$ is 0.667 with a std of 0.180. For the unaffected pixels the mean of $\hat{r}$ is 0.616 with a std of 0.180 and the mean of $\hat{r}_{13}$ is 0.699 with a std of 0.155. We notice that there is a clear difference between the estimate of r for the affected pixels and the remaining correlation estimates.

As an interesting remark we illustrate that the performance of the change detector can be improved by not only using $\hat{r}_i$ from pixel i to classify pixel i, but also using the neighbouring values $\hat{r}_j$s. This reflects the assumption of a locally homogeneous ground; the correlations among the three images for a particular pixel presumably resembles the correlations in nearby pixels. In Figure A.4(b) we have used the rule that there is a change at pixel i if

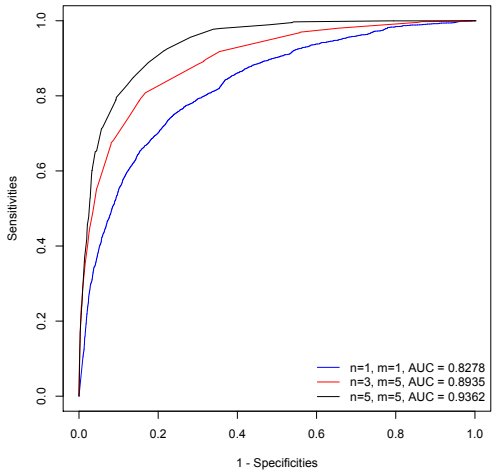
$$\#\{j : \hat{r}_j < t, j \in W_n\} \geq m, \tag{A.13}$$

where W_n is the window of size $n \times n$ centered at pixel i. This method is a new approach and as compared to Chatelain et al. (2009). We refer to the method as the neighbourhood method. We notice that if $n = m = 1$ we simply have the rule discussed above.

The three curves in Figure A.4(b) corresponds to three different classifiers based on the neighbourhood method. The values of n and m define the classifier through (A.13). For all three curves, the estimation of r is based on a window of size 3×3. As can be seen in the plot the AUC is increased when the neighbourhood method is



(a) MLE.



(b) The neighbourhood method.

Figure A.4: (a) ROCs for the classification based on MLE for three three window sizes. (b) ROCs for the classification based on the neighbourhood method. For all three classifiers the estimation of r is based on a window of size 3×3. The values of n and m refers to (A.13).

used. This implies that it is worthwhile to take the neighbourhood into account when performing the classification of a single pixel.

Acknowledgements

We would like to thank Ege Rubak for fruitful discussions and CNES Toulouse for letting us use the volcano data in section A.3.

Appendix: Saddlepoint approximation

In this appendix we state a few details on the saddlepoint approximations used in section A.3. We follow the procedure from section 2.2 in Jensen (1995). We first find the saddlepoint approximation to $p_3(n_1, n_2, n_3)$. From (A.4) we obtain the Laplace transform

$$\begin{aligned}\mathscr{L}(s) = |I - Q|^{\alpha} \\ \times \{1 - b_1 e^{s_1} - b_2 e^{s_2} - b_3 e^{s_3} - b_{12} e^{s_1+s_2} - b_{13} e^{s_1+s_3} - b_{23} e^{s_2+s_3} - b_{123} e^{s_1+s_2+s_3}\},\end{aligned}$$

where $b_i = q_{ii}$, $i = 1, 2, 3$, $b_{ij} = q_{ij}qji - q_{ii}q_{jj}$, $i < j$, and $b_{123} = |Q|$. The saddlepoint is given as the solution to the equations

$$\frac{\partial}{\partial s_i} \log \mathscr{L}(s) = n_i, \quad i = 1, 2, 3. \tag{A.14}$$

Letting $z_i = e^{s_i}$, $i = 1, 2, 3$, and solving (A.14) we find that

$$z_i = \frac{n_i(1 - b_j z_j - b_k z_k - b_{jk} z_j z_k)}{(\alpha + n_i)(b_i + b_{ij} z_j + b_{ik} z_k + b_{123} z_j z_k)}, \quad i = 1, 2, 3, \tag{A.15}$$

with $b_{ij} = b_{ji}$, $i < j$. The saddlepoint can now be found from an iterative scheme based on (A.15), and the saddlepoint approximation is then given by formula (2.2.4) in Jensen (1995).

The saddlepoint approximations for $p_3(n_1, \cdot, \cdot)$, $p_3(\cdot, n_2, \cdot)$ and $p_3(\cdot, \cdot, n_3)$ are found in a similar way. We only consider the approximation to $p_3(n_1, \cdot, \cdot)$ since the others can be found by symmetry. By differentiating (A.4) n_1 times with respect to z_1 and letting $z_1 = 0$ we obtain

$$\sum_{jk} n_1! p_3(n_1, j, k) z_2^j z_3^k = |I - Q|^{\alpha} \alpha^{\uparrow n_1} \frac{(b_1 + b_{12}z_2 + b_{13}z_3 + b_{123}z_2z_3)^{n_1}}{(1 - b_2z_2 - b_3z_3 - b_{23}z_2z_3)^{\alpha + n_1}}.$$

Normalising both sides by $\sum_{j,k} n_1! p_3(n_1, j, k)$ we get the following Laplace transform $\mathscr{L}_2$

$$\mathscr{L}_2(s_2, s_3) = \left(\frac{b_1 + b_{12}e^{s_2} + b_{13}e^{s_3} + b_{123}e^{s_2+s_3}}{b_1 + b_{12} + b_{13} + b_{123}} \right)^{n_1} \left(\frac{1 - b_2 - b_3 - b_{23}}{1 - b_2e^{s_2} - b_3e^{s_3} - b_{23}e^{s_2+s_3}} \right)^{\alpha + n_1}$$

The saddlepoint is given as the solution to the equations

$$\frac{\partial}{\partial s_i} \log \mathscr{L}_2(s_2, s_3) = n_i, \quad i = 2, 3. \tag{A.16}$$

Letting $z_i = e^{s_i}$, $i = 2, 3$, and solving (A.16) gives the two quadratic equations

$$d_i d_{1i}(\alpha + n_i) z_i^2 + (d_{i1} d_i (n_1 + n_i + \alpha) - d_{i0} d_{1i} (n_i - n_1)) z_i - n_i d_{i1} d_i = 0, \quad i = 2, 3,$$

where $d_{i0} = 1 - b_j z_j$, $d_{i1} = b_1 + b_{1j} z_j$, $d_{1i} = b_{1i} + b_{123} z_j$ and $d_i = b_i + b_{ji} z_j$, $i, j = 2, 3$, $i \neq j$, with $b_{ij} = b_{ji}$. The relevant solution to these equations is found by making sure that $z_i > 0$, $d_{i1} + d_{1i} z_i > 0$ and $d_{i0} + d_i z_i > 0$, $i = 2, 3$. Again the two equations are solved by an iterative scheme where new values are found with fixed values of the d's, and formula (2.2.4) from Jensen (1995) gives the approximation.

The saddlepoint approximation described above for the NB_3-distribution can also be considered for the NB_m-distribution, $m > 3$. However, it can only be used for large counts and the quality of the approximation may well depend on the dimension m.

Bibliography

Aitchison, J. and C. Ho (1989). The multivariate Poisson-log normal distribution. *Biometrika 76*(4), 643–653.

Chatelain, F., S. Lambert-Lacroix, and J. Tourneret (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities. *Statistics and Computing 19*(3), 283–301.

Choo, L. and S. Walker (2008). A new approach to investigating spatial variations of disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(2), 395–405.

Edwards, C. B. and J. Gurland (1961). A class of distributions applicable to accidents. *Journal of the American Statistical Association 56*(295), pp. 503–517.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Griffiths, R. (1984). Characterization of infinitely divisible multivariate gamma distributions. *Journal of Multivariate Analysis 15*(1), 13–20.

Griffiths, R. and R. K. Milne (1987). A class of infinitely divisible multivariate negative binomial distributions. *Journal of Multivariate Analysis 22*(1), 13–23.

Henderson, R. and S. Shimakura (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika 90*(2), 355–366.

Jensen, J. L. (1995). *Saddlepoint Approximations*, Volume 16 of *Oxford Statistical Science Series*. Clarendon Press Oxford.

Møller, J. and E. Rubak (2010). A model for positively correlated count variables. *International Statistical Review 78*(1), 65–80.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rubak, E., J. Møller, and P. McCullagh (2010). Statistical inference for a class of multivariate negative binomial distributions. Technical Report R-2010-10, Aalborg University.

Shirai, T. (2007). Remarks on the positivity of α-determinants. *Kyushu Journal of Mathematics 61*(1), 169–189.

Vere-Jones, D. (1997). Alpha-permanents and their applications to multivariate gamma, negative binomial and ordinary binomial distributions. *New Zealand Journal of Mathematics 26*(1), 125–149.

Supplementary material

This appendix presents applications of the above results not included in the submitted version of Paper A.

A.I Application to α-permanental random fields

The α-permanental random field (α-prf) is studied in detail in Møller and Rubak (2010). In this section we show that this model fit into the setup of Paper A and apply the results.

Let $S = \{s_1, \ldots, s_m\}$ be an arbitrary finite set, and let $N = (N_s, s \in S)$ be a collection of non-negative integer-valued random variables. The α-prf is defined as follows.

Definition A.5. *$N = \{N_s, s \in S\}$ is an α-permanental random field with parameter (α, C) if the probability generating function is given by*

$$\varphi(z) = \mathbb{E}\prod_{s\in S} z_s^{N_s} = |I + \alpha(I - Z)C|^{-1/\alpha}, \tag{A.17}$$

with $Z = \operatorname{diag}(z_1, \ldots, z_m)$, α a positive number and C a real $m \times m$ matrix.

With this definition it follows from (A.1) that the α-prf corresponds to the multivariate negative binomial model with parameters $(1/\alpha, \alpha C)$. Due to the parameterisation the formulas for the moments take a different form than (A.2). The first and second order moments of the α-prf can be found to be

$$\begin{aligned} \mathbb{E}N_i &= c_{ii}, & \mathbb{E}(N_iN_j) &= \alpha c_{ij}c_{ji} + c_{ii}c_{jj}, \quad i \neq j, \\ \operatorname{Var} N_i &= c_{ii} + \alpha c_{ii}^2, & \operatorname{Cov}(N_i, N_j) &= \alpha c_{ij}c_{ji} \quad i \neq j. \end{aligned}$$

Therefore, in the α-prf the parameter α scales the variance independently of the mean which is not the case in the multivariate negative binomial model.

The matrix Q from (A.4) becomes here $\tilde{C} = \alpha C(I + \alpha C)^{-1}$ and the reformulation of the pgf is given by

$$\varphi(z) = \left(\frac{|I - \tilde{C}|}{|I - Z\tilde{C}|} \right)^{1/\alpha}.$$

The results of the paper now applies to the α-prf; one simply has to replace a_{ij} by αc_{ij} and q_{ij} by $\tilde{c}_{ij}$.

For the rest of this section we consider the α-prf in the case $m = 3$ when C is symmetric and with identical diagonal entries. Proposition A.3 gives necessary and sufficient conditions on A for (A.1) to represent the pgf of a positive discrete infinitely divisible distribution when A is symmetric. Applying this proposition results in a simple way of characterising infinite divisibility for all $\alpha > 0$ which is the case if

$$\varphi(z) = |I + \alpha_1(I - Z)C|^{-1/\alpha_2}$$

is a pgf of a positive discrete random vector for all $\alpha_1, \alpha_2 > 0$. To deduce the conditions we apply Proposition A.3 with Q replaced by $\tilde{C}$. The matrix $\tilde{C}$ depends on α and hence the conditions should be satisfied for all $\alpha > 0$. It leads to the following corollary.

Corollary A.6. *Let $\alpha > 0$ and let*

$$C = \begin{pmatrix} d & a & b \\ a & d & e \\ b & e & d \end{pmatrix}, \tag{A.18}$$

with $d > 0$, $a, b, e \in \mathbb{R}$, satisfying $|e| \leq |a|$ and $a^2, b^2, e^2 \leq d^2$. Then the function $\varphi(z) = |I + \alpha(I - Z)C|^{-1/\alpha}$ is a pgf of a positive discrete infinitely divisible distribution for all $\alpha > 0$ if and only if one of the following conditions is satisfied.

- *At least two of the parameters a, b and e are equal to zero.*
- *$ae \neq 0$, $\operatorname{sgn}(a) = \operatorname{sgn}(e)$, $a^2, e^2 \neq d^2$ and $\frac{ae}{d} \leq b \leq \frac{ed}{a}$.*
- *$ae \neq 0$, $\operatorname{sgn}(a) \neq \operatorname{sgn}(e)$, $a^2, e^2 \neq d^2$ and $\frac{ed}{a} \leq b \leq \frac{ae}{d}$.*

The result gives rise to the following corollary that gives an alternative way of characterising infinite divisibility for all $\alpha > 0$ when C is invertible and of the type (A.18).

Corollary A.7. *Let the assumptions be as in Corollary A.6. Furthermore, assume that* $|C| \neq 0$. *The matrix C satisfies the conditions for infinite divisibility for all* $\alpha > 0$ *if and only if*

$$\operatorname{sgn}((C^{-1})_{ij}) = -\operatorname{sgn}(|C|) \cdot \operatorname{sgn}(C_{ij}), \quad i \neq j.$$

Proof. With C as defined in the corollary we obtain that

$$C^{-1} = \frac{1}{|C|} \begin{pmatrix} d^2 - e^2 & be - ad & ae - bd \\ be - ad & d^2 - b^2 & ab - ed \\ ae - bd & ab - ed & d^2 - a^2 \end{pmatrix}.$$

Then the result follows directly from Corollary A.6. □

We notice that Corollary A.7 implies that if $|C| > 0$ then the distribution is infinitely divisible if and only if the ijth element in C^{-1} is of the opposite sign as the ijth element of C when $i \neq j$. If $|C| < 0$ they should be of the same sign.

As a final remark, we notice that the setup of Henderson and Shimakura (2003) can be fit into the setup of the α-prf and therefore the results apply.

A.II Application to data: Danish testis cancer

We consider a data set from Choo and Walker (2008) on the occurrence of testis cancer in the 19 municipalities in the county of Frederiksborg, Denmark, together with the expected numbers based on population counts (see section B.4.2 for further details about the data). Rubak et al. (2010) have analysed this data using a composite likelihood based on the two-dimensional marginals with a NB_2-distribution. Here we consider the use of three-dimensional marginals following an infinitely divisible $\mathrm{NB}_3(\alpha, A)$-distribution with a symmetric matrix A with non-negative off-diagonal entries. Since the marginal means are αa_{ii} we let $\hat{a}_{ii} = E_i/\alpha$ where E_i is the expected count. The following correlation structure is assumed. For two municipalities i and j that share a border (denoted by $i \sim j$) the correlation is ρ, for two municipalities i and j that do not share a border, but share a border with the same third municipality (denoted by $i \approx j$), the correlation is σ and otherwise the correlation is zero. Let $i \equiv j$ denote that either $i \sim j$ or $i \approx j$.

The parameters ρ, σ and α are estimated by maximising a trivariate composite likelihood given by

$$\mathcal{L}_w^3(\rho, \sigma, \alpha; n) = p_2(n_J, n_S; \rho, \alpha)^{w_{JS}} \prod_{\substack{r<s<t \\ r\equiv s, r\equiv t, s\equiv t}} p_3(n_r, n_s, n_t; \rho, \sigma, \alpha)^{w_{rst}}. \tag{A.19}$$

The first term in $\mathcal{L}_w^3$ accounts for the two municipalities *Jægerspris* and *Skibby* that share a border and are separated from the other municipalities by water. It is assumed that these municipalities are not related to any other municipality. The weights of the likelihood is chosen such that when $\rho = \sigma = 0$ the composite likelihood $\mathcal{L}_w^3$ approximately coincides with the likelihood function $\mathcal{L}^1$ obtained when the observations are independent and negative binomially distributed. We

notice that $\mathcal{L}^1(\alpha;n) = \prod_{r=1}^{19} p_1(n_r;\alpha)$. For $\rho = \sigma = 0$ the composite likelihood function $\mathcal{L}_w^3$ reduces to

$$\mathcal{L}_w^3(\alpha;n) = \prod_{r=1}^{19} p_1(n_r;\alpha)^{\sum_{s\neq r, s\equiv r} w_{rs}},$$

with $w_{rs} = w_{sr}$, $r \neq s$. Hence, the weights are chosen by solving the system of linear equations given by

$$\sum_{s\neq r, s\equiv r} w_{rs} \approx 1, \quad r = 1,\ldots,19.$$

For this the procedure `lsqlin` in MATLAB was used.

Figure A.5 shows the variation in $\mathcal{L}_w^3$. For a set of (ρ,σ) the log-composite likelihood is plotted as a function of α. The log-likelihood $\log\mathcal{L}^1$ for the data, when the data is considered to be 19 independent negative binomial variables, has been added to the plot. As can be seen there is little evidence for correlation in these data. The estimate obtained by maximising (A.19) is found to be $(\hat{\alpha},\hat{\rho},\hat{\sigma}) = (36,0,0)$. Considering the case where we assume the data to be independent negative binomial variables the likelihood interval $\{\alpha|\log\mathcal{L}^1(\alpha) \geq max_{\tilde{\alpha}>0}\log\mathcal{L}^1(\tilde{\alpha}) - 2\}$ clearly covers $\alpha = \infty$. The negative binomial distribution with $\alpha = \infty$ corresponds to independent Poisson distributions.

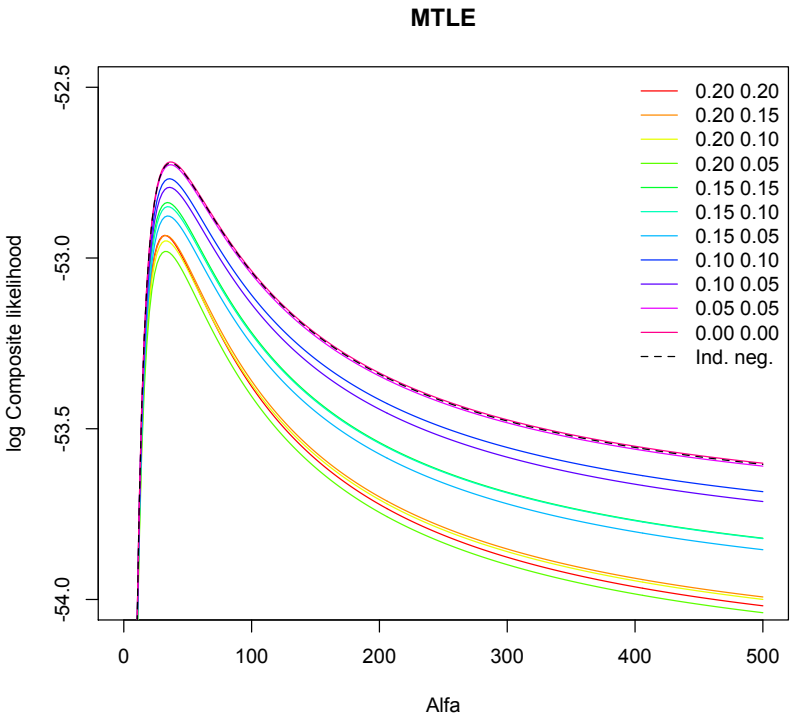


Figure A.5: Plots of $\mathcal{L}_w^3$ as functions of α for fixed values of ρ and σ. The black dashed curve represents $\mathcal{L}^1$.

Paper

B

On asymptotics results for a Poisson mixture model

Camilla Mondrup Andreassen and Jens Ledet Jensen

Department of Mathematics, Aarhus University, Denmark

Abstract: In this paper we consider a multivariate mixed Poisson model where the mixing variable arise from a function of independent and identically distributed random variables whose common distribution belongs to an exponential family. For models of this type the full likelihood is only rarely tractable and a composite likelihood based on the two-dimensional marginals is considered. The main result of the paper gives conditions for existence, consistency and asymptotic normality of the maximum pairwise likelihood estimate. The results are illustrated through a simulation study and application to data.

Keywords: Mixture models, pairwise likelihood, asymptotic distribution, Poisson distribution, gamma distribution.

B.1 Introduction

Correlated count data is often observed in applied fields and models based on mixtures of distributions are widely used to describe these data. One advantage of mixture distributions is that it is possible to construct multivariate distributions that allows for overdispersion, i.e. the variance exceeds the mean noticeably, contrary to e.g. the Poisson distribution where the variance and the mean are equal. In this paper we consider a multivariate mixed Poisson model where the mixing distribution belongs to an exponential family.

A discrete random variable Y follows a mixed Poisson distribution (Grandell, 1997) with a mixing distribution having probability density function (pdf) f, if its probability function is given by

$$\mathbb{P}(Y = y) = \int_0^\infty \frac{x^y}{y!} e^{-x} f(x)\, dx, \quad y = 0, 1, \ldots$$

This univariate case generalises naturally to the multivariate case assuming conditional independence. Let f_n be a pdf defined on $\mathbb{R}_+^n$. The corresponding probability function of the discrete random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \int_{\mathbb{R}_+^n} \prod_{l=1}^{n} \frac{x_l^{y_l}}{y_l!} e^{-x_l} f_n(\boldsymbol{x})\, d\boldsymbol{x},$$

with $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Often a log-Gaussian distribution is used as mixture distribution but the resulting distribution is quite complex. In the literature there has been several attempts to construct useful multivariate models where the mixing distribution is a multivariate gamma distribution. Henderson and Shimakura (2003) and Chatelain et al. (2009) consider a multivariate mixed Poisson distribution where the mixing distribution is a multivariate gamma distribution defined through its Laplace transform. Fiocco et al. (2009) presents a new version of the model considered by Henderson and Shimakura (2003) where the multivariate gamma distribution is constructed by use of a renewal process and the fact that the gamma distribution is infinitely divisible. We refer to Fiocco et al. (2009) for details. Another multivariate Poisson-gamma model has been presented in the paper by Choo and Walker (2008). The model is used for investigating spatial variations of disease, when the observations arise from non-infectious diseases and the investigation is concerned with spatial variations of disease risk in small areas. Their argument for the model is that two neighbouring areas share many similar features, socio-economic characteristics and levels of exposure to potential hazards for instance. Hence, it seems natural to consider the spatial correlation between neighbouring areas if one wants to create a model for disease counts. The observations are disease counts in geographical areas, and since information about the expected count in each area often is available, this is incorporated in the model. The multivariate gamma distribution used in this model is constructed by sums of independent gamma random variables. The model we present in this paper is a generalised version of the Choo and Walker (2008) model. The mixing variable considered in this paper arises from a function

of independently and identically distributed random variables with a distribution belonging to an exponential family.

A challenge of these models is that the full likelihood is only rarely tractable, and therefore makes maximum likelihood estimation complicated and often impossible. One way to overcome this problem is to consider a Bayesian approach and construct a Gibbs sampler for the posterior distribution. This is the method of Choo and Walker (2008). Another way is to consider composite likelihood estimation (see Varin et al. (2011) for an extensive overview). Since the two-dimensional probability function often is tractable, estimation based on the log-pairwise likelihood function is possible. Henderson and Shimakura (2003) and Chatelain et al. (2009) both consider maximum pairwise likelihood (MPL) estimation but only Chatelain et al. (2009) consider the asymptotic properties of the resulting estimate. Chatelain et al. (2009) consider asymptotic properties of the estimate based on N independent samples from the distribution of interest with $n = 3$ and then let N turn to infinity. In this paper we consider MPL estimation and study the asymptotic properties of the estimate in the case with one sample only, but with n turning to infinity.

The paper is organised as follows. The model is formulated in section B.2 and the asymptotic results are derived in section B.3. To conclude, a simulation study is performed and the model is applied to the data of interest in Chatelain et al. (2009). The results are given in section B.4.

B.2 Model formulation

We consider n areas (subsets) of $\mathbb{R}^2$ and define a symmetric and reflexive neighbourhood relation $\sim$ between the areas (e.g. two areas are neighbours if they share a common border). The neighbourhood A_i of area i is defined as $A_i = \{j \in \{1,\dots,n\} : j \sim i\}$. For easy reference, we let $m_i = |A_i|$ denote the number of areas in neighbourhood A_i. Furthermore, we let $A_{ij} = A_i \cup A_j$ and $m_{ij} = |A_{ij}|$.

Let Y_i be the observed count in area i and let X_i be a latent variable associated with area i. We let X be all of the latent variables and $X_L = \{X_i, i \in L\}$ for a set $L \subseteq \{1,\dots,n\}$. The counts are modelled as conditionally independent Poisson random variables

$$Y_i \,|\, X \sim \text{Po}(\beta_i \mu(X_{A_i})), \quad i = 1,\dots,n, \tag{B.1}$$

where $\beta_i \in \mathcal{B} \subseteq \mathbb{R}_+$ is a known covariate and $\mu(X_{A_i})$ is a positive function of m_i variables. The latent variables $X_1,\dots,X_n$ are assumed to be independent and identically distributed random variables with a common probability density function $f_\theta(\cdot)$ parameterised by $\theta \in \Theta \subseteq \mathbb{R}^d$, belonging to an exponential family (Choo and Walker (2008) model the X_is as independent gamma random variables and $\mu(X_{A_i}) = \sum_{l \in A_i} X_l$). That is, the density is of the form

$$f_\theta(x) = a(\theta) b(x) \exp\{\varphi(\theta) \cdot t(x)\},$$

where $\varphi : \Theta \to \mathbb{R}^k$ and $t : \mathbb{R} \to \mathbb{R}^k$ are known functions. Furthermore, it is assumed that $\varphi(\cdot)$ is C^∞. The functions $\varphi(\theta)$ and μ are chosen such that $\mathbb{E}[\mu(X_{A_i})]$ is independent of θ (and typically equal to 1). We let θ_0 denote the true value of the parameter θ and for the rest of the paper we let the expected values, variances and probabilities be with respect to θ_0 unless otherwise stated. For an example with

specific choices of f_θ and μ, see section B.4 below. We note that from the model formulation it follows directly that Y_i and Y_j are independent if $A_i \cap A_j = \emptyset$ and, furthermore, the moments of Y_i follows by use of conditional distributions.

Let $\lambda_i = \beta_i \mu(x_{A_i})$ and let $f_{\text{Po}(\lambda)}(y) = \frac{\lambda^y}{y!} e^{-\lambda}$ be the Poisson density. It follows from (B.1) that the bivariate probability function for (Y_i, Y_j), $i \neq j$, $i, j \in \mathbb{N}$, is given by

$$p_{ij}(y_i, y_j; \theta) = \int_{\mathbb{R}^{m_{ij}}} p_{ij}(y_i, y_j \,|\, x_{A_{ij}}) \prod_{l \in A_{ij}} f_\theta(x_l)\, d(x_{A_{ij}}), \tag{B.2}$$

with $p_{ij}(y_i, y_j \,|\, x_{A_{ij}}) = f_{\text{Po}(\lambda_i)}(y_i) f_{\text{Po}(\lambda_j)}(y_j)$.

For the rest of the paper we consider the following assumptions.

(B1) Every area has at least one neighbour, excluding itself, and the number of neighbours is limited, i.e. there exists $K_1 \in \mathbb{N}$, such that $2 \leq m_i \leq K_1$, $i = 1, \ldots, n$.

(B2) The set $\mathcal{B} \subseteq \mathbb{R}_+$ is compact and for any neighbourhood A, with $|A| \leq K_1$, the expected value of μ, $\mathbb{E}[\mu(X_A)]$, is finite and independent of θ.

(B3) There exist functions $c_1, c_2 > 0$ such that

(i) $\text{Var}\Big(\sum_{j \in A_i^*} \boldsymbol{u} \cdot \frac{\partial}{\partial \theta} \log p_{ij}(Y_i, 0; \theta_0) \,\Big|\, X_{A_i}\Big) \geq c_1(X_{A_i})$,

(ii) For $j \in A_i^*$: $\text{Var}\Big(\boldsymbol{u} \cdot \frac{\partial}{\partial \theta} \log p_{ij}(Y_i, 0; \theta_0) \,\Big|\, X_{A_i}\Big) \geq c_2(X_{A_i})$,

for all $i \in \mathbb{N}$ and all unit vectors $\boldsymbol{u}$.

(B4) There exists a finite constant K_2 and an open neighbourhood V_0 of θ_0 such that

$$\mathbb{E}\Big[p_{ij}(Y_i, Y_j; \theta)^{-1} \int_{\mathbb{R}^{m_{ij}}} p_3(x_{A_{ij}}) p_{ij}(Y_i, Y_j \,|\, x_{A_{ij}}) \prod_{l \in A_{ij}} f_\theta(x_l)\, d(x_{A_{ij}})\Big] \leq K_2$$

for $\theta \in V_0$ and for all polynomials p_3 in $|t_r(x_l)|$ of order less than or equal to 3, $r = 1, \ldots, k$, $l \in A_{ij}$.

B.3 Asymptotics for the maximum pairwise likelihood estimate

In this section we examine the asymptotic properties of the MPL estimate. Due to the correlation structure in the model we consider the log-pairwise likelihood function based on the bivariate probability function, p_{ij}, for the pairs (Y_i, Y_j) with $i \neq j$ and $i \sim j$. Let $A_i^* = A_i \setminus \{i\}$. The log-pairwise likelihood function is then given by

$$l_n^2(\theta) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j \in A_i^*} \log p_{ij}(Y_i, Y_j; \theta). \tag{B.3}$$

To ease the notation we let

$$U_n(\theta) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\frac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta) \quad \text{and} \quad j_n(\theta) = -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\frac{\partial^2}{\partial\theta^2}\log p_{ij}(Y_i,Y_j;\theta).$$

We notice that the number of terms in (B.3) is given by $N_n = \frac{1}{2}\sum_{i=1}^{n}|A_i^*|$. The following theorem, the main result of this paper, states that the MPL estimate of the parameter θ is consistent and asymptotic normal.

Theorem B.1. *Let $Y_1, Y_2, \ldots$ be observations from the model* (B.1) *and assume* (B1)–(B4). *Then there exists a local maximum $\hat{\theta}_n$ of $l_n^2(\theta)$ with $\hat{\theta}_n \to \theta_0$ in probability and*

$$\mathrm{Var}(U_n(\theta_0))^{-1/2}\,\mathbb{E}[j_n(\theta_0)](\hat{\theta}_n - \theta_0) \xrightarrow{\sim} N(0, I).$$

Actually, this holds for any consistent solution $\hat{\theta}_n$ to the pairwise likelihood equation.

The model and the theorem above has been formulated for $\theta \in \mathbb{R}^d$. The Cramér-Wold Theorem (Cramér and Wold, 1936) states that a Borel probability measure on $\mathbb{R}^d$ is uniquely determined by its one-dimensional projections. Using this and the uniformity with respect to $\boldsymbol{u}$ in assumption (B3), we can, without loss of generality, in the proof consider the case $d = 1$ only. Thus, we let for the rest of section B.3 the dimension of θ be $d = 1$.

The result in the theorem follows directly by Lemma B.3, B.4 and B.5 below, according to Theorem 4 in Jensen (2011a). Regarding the central limit theorem in Lemma B.3 we refer to Jensen (2011b). In that paper a setup with variables indexed by $i \in \mathbb{Z}^2$ is considered, but, following the steps of the proof, the setup of this paper is covered as well. The first variance condition of assumption (B3) is used to obtain a lower bound on the variance of U_n, and the second to get a lower bound on the mean of j_n. From Proposition B.6 below it follows that the conditions are fulfilled when μ is strictly increasing as a function of each of its arguments and $t(x)\cdot\frac{\partial}{\partial\theta}\varphi(\theta)$ is strictly monotone as a function of x when x is large. Assumption (B4) is used for the uniform convergence of $j_n(\theta)$ in Lemma B.5. The condition can be rewritten as

$$\begin{aligned}\int_{\mathbb{R}^{m_{ij}}}\left\{\sum_{y_i,y_j=0}^{\infty}\frac{p_{ij}(y_i,y_j;\theta_0)}{p_{ij}(y_i,y_j;\theta)}p_{ij}(y_i,y_j\,|\,x_{A_{ij}})\right\}p_3(x_{A_ij})\prod_{l\in A_{ij}}f_\theta(x_l)\,d(x_{A_{ij}}) \\ = \mathbb{E}_\theta\left[p_3(X_{A_{ij}})\,\mathbb{E}_\theta\left[\frac{p_{ij}(Y_i,Y_j;\theta_0)}{p_{ij}(Y_i,Y_j;\theta)}\,\middle|\,X_{A_{ij}}\right]\right].\end{aligned} \tag{B.4}$$

Hence, the assumption can be reduced to the condition that the function $\theta \mapsto \frac{p_{ij}(Y_i,Y_j;\theta_0)}{p_{ij}(Y_i,Y_j;\theta)}$ should be bounded by a function of (Y_i, Y_j) with suitable integrability properties.

Before we state and prove the three lemmas, we consider the bivariate probability function for the model (B.1) and prove an additional lemma needed for Lemma B.3, B.4 and B.5. From (B.2) it follows that for $k = 1, 2, 3$ we have that

$$\frac{\partial^k}{\partial\theta^k}p_{ij}(y_i,y_j;\theta) = \int_{\mathbb{R}^{m_{ij}}}G_\theta^k(x_{A_{ij}})p_{ij}(y_i,y_j\,|\,x_{A_{ij}})\prod_{l\in A_{ij}}f_\theta(x_l)\,d(x_{A_{ij}}), \tag{B.5}$$

with

$$\begin{aligned} G^1_\theta(x_{A_{ij}}) &= \sum_{l\in A_{ij}} g_\theta(x_l), \qquad G^2_\theta(x_{A_{ij}}) = \sum_{l\in A_{ij}} \Big\{ \tfrac{\partial}{\partial\theta} g_\theta(x_l) + g_\theta(x_l) G^1_\theta(x_{A_{ij}}) \Big\}, \\ G^3_\theta(x_{A_{ij}}) &= \sum_{l\in A_{ij}} \Big\{ \tfrac{\partial^2}{\partial\theta^2} g_\theta(x_l) + 2\tfrac{\partial}{\partial\theta} g_\theta(x_l) G^1_\theta(x_{A_{ij}}) + g_\theta(x_l) G^2_\theta(x_{A_{ij}}) \Big\}, \end{aligned} \tag{B.6}$$

where

$$g_\theta(x) = \tfrac{\partial}{\partial\theta} \log f_\theta(x) = \tfrac{\partial}{\partial\theta} \log a(\theta) + t(x) \cdot \tfrac{\partial}{\partial\theta} \varphi(\theta). \tag{B.7}$$

The following lemma ensures that the moments of $\frac{\partial^k}{\partial\theta^k} \log p_{ij}(Y_i, Y_j; \theta)$ are finite.

Lemma B.2. *Let $Y_1, Y_2, \ldots$ be observations from the model* (B.1) *and assume* (B1). *For $k = 1, 2, 3$ and $m \in \mathbb{N}$, there exists $0 < \delta_{m,k} < \infty$, such that*

$$\mathbb{E}\Big[\Big| \tfrac{\partial^k}{\partial\theta^k} \log p_{ij}(Y_i, Y_j; \theta_0) \Big|^m \Big] \le \delta_{m,k}.$$

Proof. We note that $\frac{\partial}{\partial\theta} \log p_{ij}(y_i, y_j; \theta_0) = \frac{\frac{\partial}{\partial\theta} p_{ij}(y_i, y_j; \theta_0)}{p_{ij}(y_i, y_j; \theta_0)}$ and from Minkowski's inequality it follows that

$$\begin{aligned} &\mathbb{E}\Big[\Big| \tfrac{\partial^2}{\partial\theta^2} \log p_{ij}(Y_i, Y_j; \theta_0) \Big|^m \Big] \\ &\quad \le \left(\mathbb{E}\left[\left| \frac{\frac{\partial^2}{\partial\theta^2} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)} \right|^m \right]^{1/m} + \mathbb{E}\left[\left| \frac{\frac{\partial}{\partial\theta} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)} \right|^{2m} \right]^{1/m} \right)^m \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}\Big[\Big| \tfrac{\partial^3}{\partial\theta^3} \log p_{ij}(Y_i, Y_j; \theta_0] \Big|^m \Big] \\ &\quad \le \left(\mathbb{E}\left[\left| 2\left(\frac{\frac{\partial}{\partial\theta} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)} \right)^3 - 3 \frac{\frac{\partial}{\partial\theta} p_{ij}(Y_i, Y_j; \theta_0) \frac{\partial^2}{\partial\theta^2} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)^2} \right|^m \right]^{1/m} \right. \\ &\qquad \left. + \mathbb{E}\left[\left| \frac{\frac{\partial^3}{\partial\theta^3} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)} \right|^m \right]^{1/m} \right)^m. \end{aligned} \tag{B.8}$$

Hence, in order to prove the lemma it remains to show that for $k = 1, 2, 3$ and $m \in \mathbb{N}$, there exist $0 < \delta_{m,k} < \infty$, such that

$$\mathbb{E}\left[\left| \frac{\frac{\partial^k}{\partial\theta^k} p_{ij}(Y_i, Y_j; \theta_0)}{p_{ij}(Y_i, Y_j; \theta_0)} \right|^m \right] \le \delta_{m,k}.$$

It follows from (B.2) and (B.5) that

$$\frac{\frac{\partial^k}{\partial\theta^k} p_{ij}(y_i, y_j; \theta_0)}{p_{ij}(y_i, y_j; \theta_0)} = \tilde{\mathbb{E}}[G^k_{\theta_0}(X_{A_{ij}})],$$

with $\tilde{\mathbb{E}}$ being the expected value w.r.t. the density $x_{A_{ij}} \mapsto \frac{p_{ij}(y_i,y_j|x_{A_{ij}})}{p_{ij}(y_i,y_j;\theta_0)} \prod_{k\in A_{ij}} f_{\theta_0}(x_k)$. From Jensens' inequality it now follows that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{\frac{\partial^k}{\partial\theta^k} p_{ij}(Y_i,Y_j;\theta_0)}{p_{ij}(y_i,y_j;\theta_0)}\right|^m\right] \\
&\quad\le \mathbb{E}[\tilde{\mathbb{E}}[\,|G^k_{\theta_0}(X_{A_{ij}})|^m]] \\
&\quad= \sum_{y_i,y_j=0}^{\infty} p_{ij}(y_i,y_j;\theta_0) \int_{\mathbb{R}^{m_{ij}}} \left|G^k_{\theta_0}(x_{A_{ij}})\right|^m \frac{p_{ij}(y_i,y_j|x_{A_{ij}}) \prod_{l\in A_{ij}} f_{\theta_0}(x_l)}{p_{ij}(y_i,y_j;\theta_0)}\, d(x_{A_{ij}}) \\
&\quad= \int_{\mathbb{R}^{m_{ij}}} \left|G^k_{\theta_0}(x_{A_{ij}})\right|^m \sum_{y_i,y_j=0}^{\infty} p_{ij}(y_i,y_j|x_{A_{ij}}) \prod_{l\in A_{ij}} f_{\theta_0}(x_l)\, d(x_{A_{ij}}) \\
&\quad= \int_{\mathbb{R}^{m_{ij}}} \left|G^k_{\theta_0}(x_{A_{ij}})\right|^m \prod_{l\in A_{ij}} f_{\theta_0}(x_l)\, d(x_{A_{ij}}).
\end{aligned}
$$

It follows from (B.6) that for fixed x_l, $G^k_{\theta_0}(x_{A_{ij}})^m$ is a polynomium in $t(x_l)$ with coefficients that are continuous functions of θ_0 and therefore this is finite due to properties of the exponential family. □

We are now ready to state and proof the lemmas that combined give the result of Theorem B.1.

Lemma B.3. *Let $Y_1, Y_2, \ldots$ be observations from the model* (B.1) *and assume* (B1)–(B3). *Then there exists $\varepsilon_0 > 0$, $K_0 < \infty$, such that $\varepsilon_0 \le \frac{1}{N_n}\mathrm{Var}(U_n(\theta_0)) \le K_0$ and*

$$
\frac{1}{\sqrt{\mathrm{Var}(U_n(\theta_0))}} U_n(\theta_0) \xrightarrow{\sim} N(0,1).
$$

Proof. The proof in Jensen (2011b) can be used in our setup and the result of Lemma B.3 follows if we show that there exists a finite c_0, such that for all i

$$
\mathbb{E}\Big[\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big] = 0, \quad \mathbb{E}\Big[\,\Big|\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big|^3\Big] \le c_0,
$$

and that there exists $\varepsilon_0 > 0$, such that

$$
\mathrm{Var}\left(\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\right) \ge \varepsilon_0 N_n.
$$

The first two conditions are satisfied due to standard likelihood theory and Lemma B.2. For the third, we wish to apply the law of total variance conditioning in such a way that we obtain a sum of independent variables. Consider the following recursion. Let $i_1 \in \{1,\ldots,n\}$. Then, for $k > 1$, choose $i_k \in \{1,\ldots,n\}\setminus\cup_{j=1}^{k-1} A_{i_j}$ recursively and let Ω be the largest value of k for which this is possible. Then, let $I = \cup_{j=1}^{\Omega}\{i_j\}$,

$A_\Omega = \cup_{j=1}^{\Omega} A_{i_j}^*$ and $Y^* = \{Y_j\}_{j\in A_\Omega}$. Then

$$\begin{aligned}
&\mathrm{Var}\Big(\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big)\\
&\quad\geq \mathbb{E}\Big[\mathrm{Var}\Big(\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\tfrac{\partial}{\partial\theta}\log p_{i(i+1)}(Y_i,Y_{i+1};\theta_0)\Big|\{X_l\},Y^*\Big)\Big]\\
&\quad= \sum_{i\in I}\mathbb{E}\Big[\mathrm{Var}\Big(\sum_{j\in A_i^*}\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big|\{X_l\},Y^*\Big)\Big].
\end{aligned}$$

It now suffices to show that all the terms in the sum are greater than some constant $\delta_0 > 0$. Since the calculations are identical for all terms, we only consider the first term and assume, that $i_1 = 1$. Then

$$\begin{aligned}
&\mathbb{E}\Big[\mathrm{Var}\Big(\sum_{j\in A_1^*}\tfrac{\partial}{\partial\theta}\log p_{1j}(Y_1,Y_j;\theta_0)\Big|X,Y^*\Big)\Big]\\
&\quad\geq \mathbb{E}\Big[\mathbb{P}(Y_j=0, j\in A_1^*|\{X_l\})\mathrm{Var}\Big(\sum_{j\in A_1^*}\tfrac{\partial}{\partial\theta}\log p_{1j}(Y_1,Y_j;\theta_0)|\{Y_j=0,j\in A_1^*\},X\Big)\Big]\\
&\quad= \mathbb{E}\Big[\exp\Big\{-\sum_{j\in A_1^*}\lambda_j\Big\}\mathrm{Var}\Big(\sum_{j\in A_1^*}\tfrac{\partial}{\partial\theta}\log p_{1j}(Y_1,0;\theta_0)\Big|X_{A_1}\Big)\Big]\\
&\quad\geq \mathbb{E}\Big[\exp\Big\{-\sum_{j\in A_1^*}\lambda_j\Big\}c_1(X_{A_1})\Big]. \qquad\text{(B.9)}
\end{aligned}$$

From assumption (B2) and (B3) it follows that there exists a constant $\delta_0 > 0$ such that $\mathbb{E}[\exp\{-\sum_{j\in A_1^*}\lambda_j\}c_1(X_{A_1})] \geq \delta_0$. This fulfils the proof. □

Lemma B.4. *Let $Y_1, Y_2, \ldots$ be observations from the model* (B.1) *and assume* (B1)–(B3). *Then*

1. $\mathbb{E}\Big[\frac{1}{N_n}j_n(\theta_0)\Big] \geq c_0$, *for some* $c_0 > 0$, *and*
2. $\frac{1}{N_n}j_n(\theta_0) - \mathbb{E}\Big[\frac{1}{N_n}j_n(\theta_0)\Big] \to 0$ *in probability.*

Proof. For the first part we notice that

$$\mathbb{E}[j_n(\theta_0)] = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\mathrm{Var}\Big(\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big).$$

Therefore, it is sufficient to show that there exists a $\delta_0 > 0$ such that

$$\mathrm{Var}\Big(\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i,Y_j;\theta_0)\Big) \geq \delta_0$$

for all $j \in A_i^*$, $i = 1, \ldots, n$. Using the same type of arguments as in (B.9) we obtain that

$$\begin{aligned}
&\mathrm{Var}\Big(\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i, Y_j; \theta_0)\Big) \\
&\quad \geq \mathbb{E}\Big[\mathbb{P}(Y_j = 0 \,|\, X_{A_i})\,\mathrm{Var}\Big(\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i, Y_j; \theta_0) \,|\, Y_j = 0, X_{A_{ij}}\Big)\Big] \\
&\quad = \mathbb{E}\Big[e^{-\lambda_j}\,\mathrm{Var}\Big(\tfrac{\partial}{\partial\theta}\log p_{ij}(Y_i, 0; \theta_0) \,|\, X_{A_i}\Big)\Big] \\
&\quad \geq \mathbb{E}[e^{-\lambda_j} c_2(X_{A_i})].
\end{aligned}$$

From assumption (B2) and (B3) it follows that there exists a constant $\delta_0 > 0$ such that $\mathbb{E}[e^{-\lambda_j} c_2(X_{A_i})] \geq \delta_0$.

For the second part it follows from Chebychev's inequality that it suffices to show that

$$\mathrm{Var}\left(\tfrac{1}{2N_n}\sum_{i=1}^{n}\sum_{j\in A_i^*}\tfrac{\partial^2}{\partial\theta^2}\log p_{ij}(Y_i, Y_j; \theta_0)\right) \to 0, \quad n \to \infty.$$

Let $\bar{A}_{ij} = \cup_{u\in A_{ij}} A_u$. We note that due to the Cauchy-Schwartz inequality and Lemma B.2, there exists a constant $0 < M < \infty$ such that

$$\Big|\mathrm{Cov}\Big(\tfrac{\partial^2}{\partial\theta^2}\log p_{ij}(Y_i, Y_j; \theta_0), \tfrac{\partial^2}{\partial\theta^2}\log p_{lm}(Y_l, Y_m; \theta_0)\Big)\Big| \leq M,$$

for $j \in A_i^*$, $m \in A_l^*$ and $i, l = 1 \ldots, n$. Due to assumption (B1) it now follows that

$$\begin{aligned}
&\mathrm{Var}\left(\tfrac{1}{2N_n}\sum_{i=1}^{n}\sum_{j\in A_i^*}\tfrac{\partial^2}{\partial\theta^2}\log p_{ij}(Y_i, Y_j; \theta_0)\right) \\
&\quad \leq \frac{1}{4N_n^2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\sum_{l=1}^{n}\sum_{m\in A_l^*}\Big|\mathrm{Cov}\Big(\tfrac{\partial^2}{\partial\theta^2}\log p_{ij}(Y_i, Y_j; \theta_0), \tfrac{\partial^2}{\partial\theta^2}\log p_{lm}(Y_l, Y_m; \theta_0)\Big)\Big| \\
&\quad \leq \frac{1}{2N_n^2}\sum_{i=1}^{n}\sum_{j\in A_i^*}\sum_{l\in\bar{A}_{ij}}\sum_{m\in A_l^*} M \\
&\quad \leq \frac{M}{N_n^2} n K_1^4.
\end{aligned}$$

This fulfils the proof since N_n^2 is of order n^2. □

Lemma B.5. *Let $Y_1, Y_2, \ldots$ be observations from the model* (B.1) *and assume* (B1)–(B4). *Then for all $\{\delta_n\}$, with $\delta_n \to 0$ for $n \to \infty$*

$$\sup_{\{\theta : |\theta - \theta| \leq \delta_n\}} |\tfrac{1}{N_n}(j_n(\theta) - j_n(\theta_0))| \xrightarrow{P} 0.$$

Proof. Let $\{\delta_n\}$ be a sequence of positive real numbers, with $\delta_n \to 0$ for $n \to \infty$. Let $B_0(\delta_n)$ be an open ball with radius δ_n centered at θ_0. Let $\theta \in B_0(\delta_n)$. Then

$$\begin{aligned}
&\sup_{\theta\in B_0(\delta_n)} |\tfrac{1}{N_n}(j_n(\theta) - j_n(\theta_0))| \\
&\quad\le \sup_{\theta\in B_0(\delta_n)} \tfrac{1}{2N_n} \sum_{i=1}^{n} \sum_{j\in A_i^*} \left| \tfrac{\partial^2}{\partial\theta^2} \log p_{ij}(Y_i, Y_j; \theta) - \tfrac{\partial^2}{\partial\theta^2} \log p_{ij}(Y_i, Y_j; \theta_0) \right| \\
&\quad= \tfrac{1}{2N_n} \sum_{i=1}^{n} \sum_{j\in A_i^*} \int_{\theta_0-\delta_n}^{\theta_0+\delta_n} \left| \tfrac{\partial^3}{\partial\theta^3} \log p_{ij}(Y_i, Y_j; \nu) \right| d\nu.
\end{aligned}$$

Let $\varepsilon > 0$. It follows from Markov's inequality that

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\theta\in B_0(\delta_n)} |\tfrac{1}{N_n}(j_n(\theta) - j_n(\theta_0))| \ge \varepsilon \Big) \\
&\quad\le \mathbb{P}\left(\tfrac{1}{2N_n} \sum_{i=1}^{n} \sum_{j\in A_i^*} \int_{\theta_0-\delta_n}^{\theta_0+\delta_n} \left| \tfrac{\partial^3}{\partial\theta^3} \log p_{ij}(Y_i, Y_j; \nu) \right| d\nu \ge \varepsilon \right) \\
&\quad\le \tfrac{1}{2\varepsilon N_n} \sum_{i=1}^{n} \sum_{j\in A_i^*} \int_{\theta_0-\delta_n}^{\theta_0+\delta_n} \mathbb{E}\Big[\left| \tfrac{\partial^3}{\partial\theta^3} \log p_{ij}(Y_i, Y_j; \nu) \right| \Big] d\nu.
\end{aligned}$$

From (B.8) it follows that

$$\begin{aligned}
\mathbb{E}\Big[\left| \tfrac{\partial^3}{\partial\theta^3} \log p_{ij}(Y_i, Y_j; \theta) \right| \Big] &\le \mathbb{E}\left[\left| \frac{\frac{\partial^3}{\partial\theta^3} p_{ij}(Y_i, Y_j; \theta)}{p_{ij}(Y_i, Y_j; \theta)} \right| \right] + 2\,\mathbb{E}\left[\left| \left(\frac{\frac{\partial}{\partial\theta} p_{ij}(Y_i, Y_j; \theta)}{p_{ij}(Y_i, Y_j; \theta)} \right) \right|^3 \right] \\
&\quad + 3\,\mathbb{E}\left[\left| \frac{\frac{\partial^2}{\partial\theta^2} p_{ij}(Y_i, Y_j; \theta)}{p_{ij}(Y_i, Y_j; \theta)} \right|^{3/2} \right]^{2/3} \mathbb{E}\left[\left| \frac{\frac{\partial}{\partial\theta} p_{ij}(Y_i, Y_j; \theta)}{p_{ij}(Y_i, Y_j; \theta)} \right|^3 \right]^{1/3}.
\end{aligned}$$

As in the proof of Lemma B.2 it follows that

$$\mathbb{E}\left[\left| \frac{\frac{\partial^k}{\partial\theta^k} p_{ij}(Y_i, Y_j; \theta)}{p_{ij}(Y_i, Y_j; \theta)} \right|^m \right] \le \mathbb{E}\left[\frac{\int_{\mathbb{R}^{m_{ij}}} |G_\theta^k(x_{A_{ij}})|^m p_{ij}(Y_i, Y_j \,|\, x_{A_{ij}}) \prod_{l\in A_{ij}} f_\theta(x_l)\, d(x_{A_{ij}})}{p_{ij}(Y_i, Y_j; \theta)} \right]$$

with $mk = 3$ for $k = 1, 2, 3$. The result now follows from assumption (B4). □

The next proposition gives sufficient conditions for the conditions in assumption (B3) to be fulfilled.

Proposition B.6. *Assumption* (B3) *is fulfilled if the functions* $x_l \mapsto \mu(x_{A_i})$, $l \in A_i$, $i = 1, \ldots, n$, *are strictly increasing and if there exists a constant* $c \in \mathbb{R}$ *such that the function* $x \mapsto t(x) \cdot \frac{\partial}{\partial\theta}\varphi(\theta)$ *is strictly increasing (or decreasing) for* $x > c$.

Proof. We consider only the first condition of assumption (B4) since both conditions follow from the same type of argument. By use of (B.2)–(B.7) we obtain

that

$$\begin{aligned}\sum_{j\in A_i^*}\frac{\partial}{\partial\theta}\log p_{ij}(y_i,0;\theta_0) &= \sum_{j\in A_i^*}\frac{\int_{\mathbb{R}^{m_{ij}}} G^1_{\theta_0}(x_{A_{ij}})p_{ij}(y_i,0\,|\,x_{A_{ij}})\prod_{l\in A_{ij}} f_{\theta_0}(x_l)\,d(x_{A_{ij}})}{\int_{\mathbb{R}^{m_{ij}}} p_{ij}(y_i,0\,|\,x_{A_{ij}})\prod_{l\in A_{ij}} f_{\theta_0}(x_l)\,d(x_{A_{ij}})}\\ &= \sum_{j\in A_i^*}\frac{\int_{\mathbb{R}^{m_{ij}}} G^1_{\theta_0}(x_{A_{ij}})e^{y_i\log\lambda_i-\lambda_i-\lambda_j}\prod_{l\in A_{ij}} f_{\theta_0}(x_l)\,d(x_{A_{ij}})}{\int_{\mathbb{R}^{m_{ij}}} e^{y_i\log\lambda_i-\lambda_i-\lambda_j}\prod_{l\in A_{ij}} f_{\theta_0}(x_l)\,d(x_{A_{ij}})}\\ &= \sum_{j\in A_i^*}\mathbb{E}_{y_i}[G^1_{\theta_0}(X_{A_{ij}})],\end{aligned}$$

with $\mathbb{E}_{y_i}$ being the expected value with respect to the density

$$x_{A_{ij}}\mapsto\frac{e^{y_i\log\lambda_i-\lambda_i-\lambda_j}\prod_{l\in A_{ij}} f_{\theta_0}(x_l)}{\int_{\mathbb{R}^{m_{ij}}} e^{y_i\log\lambda_i-\lambda_i-\lambda_j}\prod_{l\in A_{ij}} f_{\theta_0}(x_l)\,d(x_{A_{ij}})}.$$

When y_i increases the term $e^{y_i\log\lambda_i-\lambda_i}$ shifts the mode to larger values of λ_i and hence larger values of the x_ls since the functions $x_l\mapsto\lambda_i=\beta_i\mu(x_{A_i})$ are strictly increasing. We notice that

$$G^1_{\theta_0}(x_{A_{ij}}) = m_{ij}\frac{\partial}{\partial\theta}\log a(\theta_0)+\sum_{l\in A_{ij}} t(x_l)\cdot\frac{\partial}{\partial\theta}\varphi(\theta_0).$$

By assumption we then have that $G^1_{\theta_0}$ is strictly increasing (or decreasing) for large values of the x_ls. This implies that there exists a constant $0<c<\infty$ such that $y_i\mapsto\sum_{j\in A_i^*}\frac{\partial}{\partial\theta}\log p_{ij}(y_i,0;\theta_0)$ is strictly increasing (or decreasing) for $y_i>c$. We can therefore conclude that $\mathrm{Var}\big(\sum_{j\in A_i^*}\frac{\partial}{\partial\theta}\log p_{ij}(Y_i,0;\theta_0)\,|\,X_{A_i}\big)>0$. □

We note that the assumption about Poisson distribution is not essential. The only place the Poisson distribution plays a role is in the proof of Proposition B.6. Hence, the result in Theorem B.1 can be generalised to other positive discrete distributions.

In section B.4 below we consider the setup with f_θ being the density function of a gamma distribution with mean 1 and variance θ^{-1}. Before we turn to that setup, we notice that the conditions of Proposition B.6 are fulfilled for the classical Poisson log-normal setup (Aitchison and Ho, 1989) with f_θ being the density of a normal distribution with mean $-\theta/2$ and variance θ. Mathematically, $f_\theta(x)=\frac{1}{\sqrt{2\pi\theta}}\exp\{-\frac{(x+\theta/2)^2}{2\theta}\}$ for $x\in\mathbb{R}$ and $\mu(x_{A_i})=\exp\{\sum_{l\in A_i}x_l\}$, $i=1,\ldots,n$ (this ensures that $\mu>0$ and $\mathbb{E}[\mu(X_{A_i})]=1$, $i=1,\ldots,n$). It follows that μ is strictly increasing in all of its arguments and that $t(x)\cdot\frac{\partial}{\partial\theta}\varphi(\theta)=\frac{1}{2\theta^2}x^2$ is strictly increasing for $x>0$.

B.4 A Poisson-gamma setup

To illustrate the main result of the paper we consider the following example where the latent variables are modelled as independent gamma random variables with mean one (Choo and Walker, 2008). For the rest of this section we let

$$f_\theta(x)=\frac{\theta^\theta}{\Gamma(\theta)}e^{(\theta-1)\log x-\theta x},\quad x>0,\,\theta>0, \tag{B.10}$$

and as 'correlation'-function μ we let $\mu(x_{A_i}) = \frac{1}{m_i}\sum_{j\in A_i} x_j$ implying that assumption (B2) is fulfilled with $\mathbb{E}[\mu(X_{A_i})] = 1$, $i = 1,\dots,n$. It follows from Proposition B.6 that assumption (B3) is fulfilled for this setup since $t(x)\cdot\frac{\partial}{\partial\theta}\varphi(\theta) = \log x - x$, which is increasing for $x > 1$ and μ is strictly increasing in all of its arguments. Let $m^0_{ij} = |A_i \cap A_j|$. From the model formulation it then follows that

$$\begin{aligned}\operatorname{Cor}(Y_i, Y_j) &= \frac{\beta_i\beta_j \operatorname{Cov}(\mu(X_{A_i}), \mu(X_{A_j}))}{\sqrt{(\beta_i \mathbb{E}[\mu(X_{A_i})] + \beta_i^2 \operatorname{Var}(\mu(X_{A_i})))(\beta_j \mathbb{E}[\mu(X_{A_j})] + \beta_j^2 \operatorname{Var}(\mu(X_{A_j})))}} \\ &= \frac{m^0_{ij}}{\sqrt{m_i m_j}} \frac{1}{\sqrt{(1+\theta m_i/\beta_i)(1+\theta m_j/\beta_j)}}, \end{aligned} \tag{B.11}$$

for $i \neq j$ with $A_i \cap A_j \neq \emptyset$. Therefore, with this definition of μ, large values of θ correspond to weak dependence between the counts.

B.4.1 Simulation

We perform a simulation study to illustrate the main result of the paper. For simplicity we consider the one-dimensional case, that is, instead of subsets of $\mathbb{R}^2$ one can think of the areas as consecutive intervals of the real line, $\mathbb{R}$. Hence, the neighbourhood of interval i becomes $A_i = \{i-1, i, i+1\}$ with $m_i = 3$, $i = 1,\dots,n$, implying that assumption (B1) is fulfilled. With this setup the log-pairwise likelihood reduces to

$$l^2_n(\theta) = \sum_{i=1}^{n-1} \log p_{i,i+1}(Y_i, Y_{i+1}; \theta).$$

Furthermore, we are able to derive a closed form of the bivariate probability function of interest by use of (B.2) and the binomial theorem. For $u, v \in \mathbb{N}_0$ it is given by

$$\begin{aligned} &p_{i,i+1}(u,v;\theta) \\ &\quad= \frac{(\beta_i/m_i)^u}{u!}\frac{(\beta_{i+1}/m_{i+1})^v}{v!}\frac{\theta^{4\theta}}{\Gamma(2\theta)\Gamma(\theta)^2} \\ &\quad\times \sum_{l=0}^{u}\sum_{j=0}^{v}\binom{u}{l}\binom{v}{j}\frac{\Gamma(l+j+2\theta)}{\left(\frac{\beta_i}{m_i}+\frac{\beta_{i+1}}{m_{i+1}}+\theta\right)^{l+j+2\theta}}\frac{\Gamma(u-l+\theta)}{\left(\frac{\beta_i}{m_i}+\theta\right)^{u-l+\theta}}\frac{\Gamma(v-j+\theta)}{\left(\frac{\beta_{i+1}}{m_{i+1}}+\theta\right)^{v-j+\theta}}. \end{aligned} \tag{B.12}$$

From this we can show that assumption (B4) is fulfilled. Let $0 \le \delta < 1$ and let $\theta_\delta = \theta \pm \delta$. Furthermore, let $a \in \mathbb{R}_+$ and $k \in \mathbb{N}_0$ and consider the following inequalities.

$$\frac{(a+\theta)^{k+\theta}}{(a+\theta_\delta)^{k+\theta_\delta}} \le c(\theta,\delta)\left(\frac{k+\theta}{k+\theta-\delta}\right)^{k+\theta} \le c(\theta,\delta)(1+\varepsilon(\theta,\delta))^k$$

and

$$\frac{\Gamma(k+\theta)}{\Gamma(k+\theta_\delta)} \le e^{|\log\Gamma(k+\theta)-\log\Gamma(k+\theta_\delta)|} \le c(\theta,\delta)e^{\varepsilon(\theta,\delta)\log(1+k)},$$

with $c(\theta,\delta)$ and $\varepsilon(\theta,\delta)$ being generic positive and finite constants dependent on θ and δ. Moreover, $\varepsilon(\theta,\delta)\to 0$ for $\delta\to 0$. Using each of these three times we obtain

$$\begin{aligned}\frac{p_{i,i+1}(u,v;\theta\pm\delta)}{p_{i,i+1}(u,v;\theta)} &\le c(\theta,\delta)(u^4+v^4+1)(1+\varepsilon(\theta,\delta))^{u+v}\\ &\le \tilde{c}(\theta,\delta)(1+\varepsilon(\theta,\delta))^{u+v},\end{aligned}$$

where $\tilde{c}(\theta,\delta)$ is a generic positive and finite constants dependent on θ and δ. Considering the expression of (B.4) we obtain that

$$\begin{aligned}&\mathbb{E}_\theta\Big[p_3(X_{A_{i,i+1}})\,\mathbb{E}_\theta\Big[\frac{p_{i,i+1}(Y_i,Y_{i+1};\theta\pm\delta)}{p_{i,i+1}(Y_i,Y_{i+1};\theta)}\,\Big|\,X_{A_{i,i+1}}\Big]\Big]\\ &\quad\le c(\theta,\delta)\,\mathbb{E}_\theta\Big[p_3(X_{A_{i,i+1}})\,\mathbb{E}_\theta\Big[(1+\varepsilon(\theta,\delta))^{Y_i+Y_{i+1}}\,\Big|\,X_{A_{i,i+1}}\Big]\Big]\\ &\quad= c(\theta,\delta)\,\mathbb{E}_\theta\Big[p_3(X_{A_{i,1+i}})e^{\varepsilon(\theta,\delta)\beta_i\mu(X_{A_i})}e^{\varepsilon(\theta,\delta)\beta_{i+1}\mu(X_{A_{i+1}})}\Big].\end{aligned}$$

From (B.10) it follows that this is finite since $\beta_i\varepsilon(\theta,\delta)<\theta$ for δ sufficiently small.

For comparison in the numerical results, we also estimate θ by the method of moments. We notice that $\mathbb{E}[Y_i/\beta_i]=1$ and

$$\operatorname{Cov}\Big(\tfrac{Y_i}{\beta_i},\tfrac{Y_j}{\beta_j}\Big)=\operatorname{Cov}(\mu(X_{A_i}),\mu(X_{A_j}))=\begin{cases}\frac{2}{9\theta} & \text{for } |i-j|=1\\ \frac{1}{9\theta} & \text{for } |i-j|=2\\ 0 & \text{otherwise.}\end{cases}$$

An estimate of $\varphi=1/\theta$ can then be found as

$$\hat{\varphi}_{\mathrm{MOM}}=\tfrac{9}{2n-3}\Big\{\tfrac{1}{2}\sum_{i=1}^{n-1}\Big(\tfrac{y_i}{\beta_i}-1\Big)\Big(\tfrac{y_{i+1}}{\beta_{i+1}}-1\Big)+\sum_{i=1}^{n-2}\Big(\tfrac{y_i}{\beta_i}-1\Big)\Big(\tfrac{y_{i+2}}{\beta_{i+2}}-1\Big)\Big\}.$$

To create the sequence $\{\beta_i\}$ to be used in the simulation, we simulated independent observations from a $N(15,81)$-distribution and then uniformly sampled 5000 observations from the interval $[6,45]$. The true value of the parameter is $\theta=2$.

Table B.1 and Figure B.1 show the results of the simulation. The empirical bias (bias), standard deviation (std) and root mean squared error (rmse) of the logarithm of the estimated parameter θ has been reported in Table B.1. Q-Q plots for the logarithm of the MPL estimate is shown in Figure B.1. The number of Monte Carlo runs is 1000. As expected from Theorem B.1, the bias, std and rmse decreases as n increases and the Q-Q plots show that the asymptotic distribution is normal. Furthermore, we notice that MPL performs better than the method of moments.

B.4.2 Application to data

We consider the data set from Choo and Walker (2008) on the occurence of testis cancer in the 19 municipalities in the county of Frederiksborg, Denmark, together with the expected numbers based on population counts. Due to the definition of f_θ and μ it is possible to find a closed form of the bivariate probability function of interest which resembles (B.12). Two municipalities are neighbours if and only if they share a common border. Hence, the A_is have been defined by considering a

Table B.1: Bias, std and rmse of the logarithm of the estimate.

	MPL			Moment		
n	bias	std	rmse	bias	std	rmse
20	0.1380	0.6045	0.6197	0.5374	1.1394	1.2592
50	0.0568	0.3614	0.3657	0.2451	0.7696	0.8073
100	0.0332	0.2491	0.2512	0.1342	0.5404	0.5565
250	0.0119	0.1521	0.1524	0.0332	0.2751	0.2769
500	0.0098	0.1106	0.1110	0.0306	0.2039	0.206
1000	0.0072	0.0771	0.0774	0.0155	0.1395	0.1402
5000	0.0005	0.0351	0.0351	0.0005	0.0630	0.0630

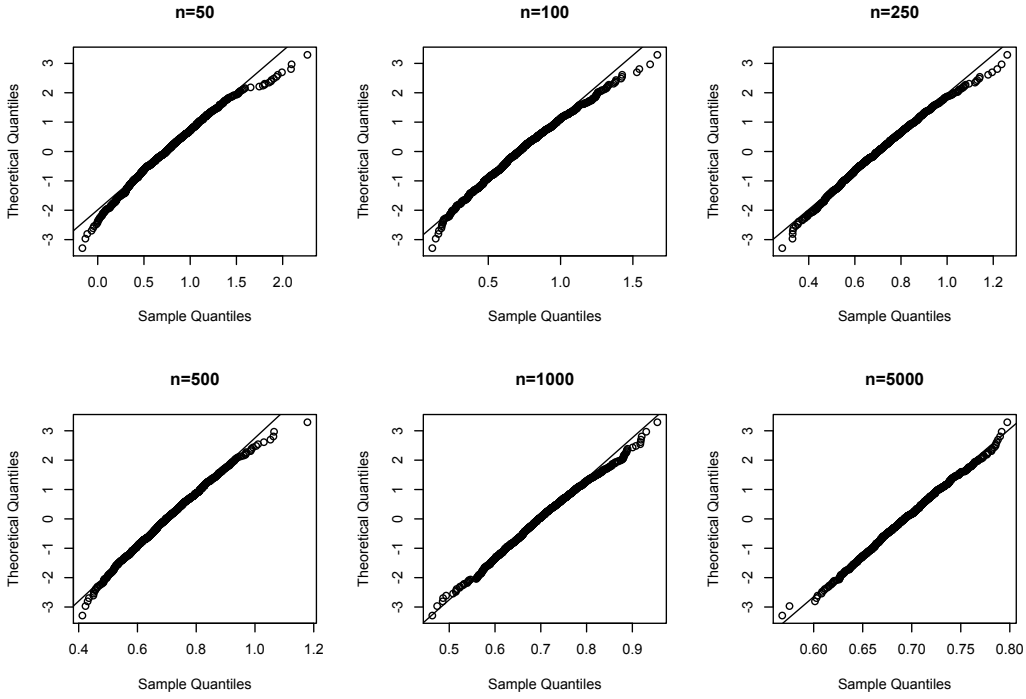


Figure B.1: Normal Q-Q plots of the logarithm of the estimated parameter θ.

Table B.2: Danish testis cancer data.

i	Municipality name	Y_i	β_i	A_i
1	Allerød	18	17.61	{1,2,3,10,12,14,17,18}
2	Birkerød	17	18.20	{1,2,3,12}
3	Farum	14	13.65	{1,2,3,18}
4	Fredensborg-Humlebæk	14	14.29	{4,7,9,10,14}
5	Frederikssund	21	13.17	{5,6,16,17,19}
6	Frederiksværk	14	14.63	{5,6,8,11,16}
7	Græsted-Gilleleje	13	12.38	{4,7,8,9,10}
8	Helsinge	8	13.66	{6,7,8,10}
9	Helsingør	31	47.18	{4,7,9}
10	Hillerød	28	27.23	{1,4,7,8,10,14,16,17}
11	Hundested	8	6.44	{6,11}
12	Hørsholm	28	17.04	{1,2,12,14}
13	Jægerspris	4	6.05	{13,15}
14	Karlebo	12	13.78	{1,4,10,12,14}
15	Skibby	6	4.57	{13,15}
16	Skævinge	6	4.28	{5,6,10,16,17}
17	Slangerup	3	6.44	{1,5,10,16,17,18,19}
18	Stenløse	13	10.47	{1,3,17,18,19}
19	Ølstykke	14	10.93	{5,17,18,19}

geographic map. As covariate information we let $\beta_i = E_i$, $i = 1,\dots,n$, where E_i is the expected count in municipality i. The data is shown in Table B.2.

Estimating the parameter θ by MPL gives the result $\hat{\theta} = 16.93$. A histogram of the resulting correlation estimates, calculated by use of (B.11), can be seen in Figure B.2. A simulation under $\theta_0 = \infty$ (corresponding to the model where the Y_is are independent and $Y_i \sim \text{Po}(\beta_i)$, $i = 1,\dots,n$) shows that a value of $\hat{\theta}$ less than 16.93 happens with probability 0.211. The result coincides with the conclusion of Rubak et al. (2010). In that paper, they consider another type of model where spatial correlation also is taken into account, and reach the conclusion that there is little evidence of spatial correlation in the data.

Bibliography

Aitchison, J. and C. Ho (1989). The multivariate Poisson-log normal distribution. *Biometrika 76*(4), 643–653.

Chatelain, F., S. Lambert-Lacroix, and J. Tourneret (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities. *Statistics and Computing 19*(3), 283–301.

Choo, L. and S. Walker (2008). A new approach to investigating spatial variations of disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(2), 395–405.

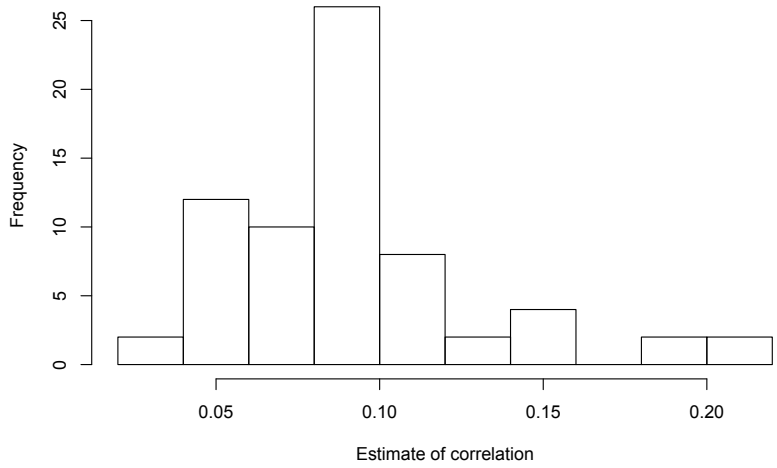


Figure B.2: Histogram of the estimated correlations between neighbouring municipalities.

Cramér, H. and H. Wold (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society 1*(4), 290–294.

Fiocco, M., H. Putter, and J. Van Houwelingen (2009). A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics 10*(2), 245–257.

Grandell, J. (1997). *Mixed Poisson Process*, Volume 77 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Henderson, R. and S. Shimakura (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika 90*(2), 355–366.

Jensen, J. L. (2011a). Asymptotic normality of m-estimators in nonhomogeneous hidden markov models. *Journal of Applied Probability 48A*, 295–306.

Jensen, J. L. (2011b). Central limit theorem for functions of weakly dependent variables. In *Proceeding of ISI-meeting in Dublin, 2011*. ISI.

Karlis, D. and E. Xekalaki (2005). Mixed Poisson distributions. *International Statistical Review 73*(1), 35–58.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rubak, E., J. Møller, and P. McCullagh (2010). Statistical inference for a class of multivariate negative binomial distributions. Technical Report R-2010-10, Aalborg University.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Paper

C

On bivariate time series of counts

Camilla Mondrup Andreassen[1], Jens Ledet Jensen[1] and Richard A. Davis[2]

[1] *Department of Mathematics, Aarhus University, Denmark*
[2] *Department of Statistics, Columbia University, New York*

Abstract: The focus of this paper is on modelling bivariate time series of counts. We consider a Poisson-based bivariate INGARCH model where the bivariate Poisson distribution is constructed by trivariate reduction of independent Poisson variables. We show that stability results previously obtained in the literature extend to the case where the Poisson distribution is replaced by a general distribution from the exponential family. Furthermore, we show that the maximum likelihood estimator of the parameter of the model is strongly consistent. A limitation of the model is that it is not able to capture negative dependence between the two time series. We propose a new bivariate Poisson distribution to replace the distribution in the bivariate INGARCH model. It is constructed by use of copulas, which allows it to capture negative dependence between the two time series at a given point in time. We show that the aforementioned stability result also applies in this new setup. The two types of models are compared through a simulation study, and both models are applied to a real data example. This is work in progress.

Keywords: Copula, time series, likelihood inference, asymptotics, count data.

C.1 Introduction

Multivariate count data are frequently encountered in applied fields and the data are often observed over time, resulting in multivariate time series of counts. The focus of this paper is on modelling bivariate time series of counts. When constructing a model for bivariate (or multivariate) time series both the modelling of the dependence between the two time series at a given point in time and the modelling of the serial dependence needs to be considered. For the univariate case, the serial dependence is often modelled by use of an autoregressive (AR), autoregressive conditional heteroskedasticity (ARCH) or generalised ARCH (GARCH) model. It is not a simple matter to model the dependence between the two time series at a given point in time (conditional on previous events) as there do not exist natural generalisations to two dimensions for many of the standard discrete distributions (e.g. the Poisson and the negative binomial distributions). For a comprehensive treatment of bivariate (and multivariate) Poisson distributions we refer to Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997).

There are several models in the literature that describe univariate time series of counts, see for instance Davis et al. (1999), McKenzie (2003) and Jung and Tremayne (2006). The treatment of the bivariate case is not as extensive but some models can be found in for instance Heinen and Rengifo (2007), Pedeli and Karlis (2011) and Liu (2012, Chapter 4). The model proposed in Liu (2012) is a Poisson-based bivariate integer-valued GARCH (INGARCH) model which is capable of capturing the serial dependence between two time series of counts. As the bivariate Poisson distribution of that model is constructed by using trivariate reduction it can not capture negative dependence between the two time series, which is a limitation.

In this paper we propose a new model for bivariate time series of counts that is able to capture negative dependence between the time series under certain assumptions. The model is a Poisson-based bivariate INGARCH model of the same type as the model proposed in Liu (2012), i.e. the structure of the conditional mean process is the same, and the marginal counts are conditionally Poisson distributed. The difference lies in the modelling of the dependence between the two time series at a given time. In the new model the dependence is modelled by a copula, i.e. the bivariate Poisson distribution, is constructed by combining two univariate Poisson distributions by a copula.

We use copulas, which, loosely speaking, are functions that couple marginal distribution functions. In other words, a cumulative distribution function (cdf) can be written in terms of marginal distribution functions and a copula, where the marginal distribution functions describe the distribution of the marginals, and the copula describes the dependence between the marginals. Mathematically, a function $C : [0,1]^d \to [0,1]$ is a d-dimensional copula, if C is a cdf of a d-dimensional random vector on $[0,1]^d$ with uniform marginals (Joe, 1997; Nelsen, 2006). From Sklar's theorem (Sklar, 1959) it follows that if C is a copula and F and G are distribution functions, then the function $H(x,y) = C(F(x), G(y))$ is a joint distribution function with marginals F and G. The reverse is also true, but the copula C is only uniquely determined if F and G are continuous. Since Sklar's theorem also holds in d dimensions, the results of this paper can presumably be extended to d dimensions.

We consider Archimedean copulas, since these possess two great properties compared to, for instance, non-parametric copulas. Firstly, Archimedean copulas allow modelling of dependence in arbitrarily high dimensions by use of only one parameter. Secondly, the most common Archimedean copulas (e.g. the Frank and the Clayton copulas) have an explicit formula for the cdf. Archimedean copulas are copulas of the form $C(u,v) = \psi^{[-1]}(\psi(u)+\psi(v))$, where the generator $\psi : [0,1]^2 \to [0,\infty]$ is a strictly decreasing, convex function with $\psi(1) = 0$, and $\psi^{[-1]}$ is the pseudo-inverse satisfying $\psi(\psi^{[-1]}(t)) = \min(t, \psi(0))$. Details about the Frank and Clayton copulas can be found in section C.4.1.

The paper is organised as follows. In section C.2, the common structure of the model presented in Liu (2012) and our new model is described. Section C.3 gives further details of the model in Liu (2012), and an extension of it where the Poisson distribution is replaced by an exponential family is considered. Furthermore, consistency of the maximum likelihood estimator is proved for the model in Liu (2012). Details of our new model are given in section C.4, and a stability result for the conditional mean process is proved. In section C.5, a simulation study is performed, and in section C.6 the two models are applied to a real data example.

C.2 Structure of the models

In this section, we consider the common structure of the model presented in Liu (2012) and our new model but defer the details to the following two sections. The basic structure of both models is a Poisson-based bivariate INGARCH model which is able to capture the serial dependence between two time series of counts.

Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2})^T$ be the bivariate observation at time t, where $\{Y_{t,1}, t \geq 1\}$ and $\{Y_{t,2}, t \geq 1\}$ are the two time series of interest. A Poisson-based bivariate INGARCH model of order $(1,1)$ is defined as

$$\mathbf{Y}_t \,|\, \mathcal{F}_{t-1} \sim \mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi), \quad \boldsymbol{\lambda}_t = (\lambda_{t,1}, \lambda_{t,2})^T = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}, \tag{C.1}$$

where $\mathcal{F}_t = \sigma\{\boldsymbol{\lambda}_1, \mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$ is the σ-algebra of past events, $\varphi \in I_\varphi$ where I_φ is a subset of $\mathbb{R}$, $\boldsymbol{\delta} = (\delta_1, \delta_2) \in \mathbb{R}_+^2$ and $\mathbf{A}, \mathbf{B}$ are both 2×2 matrices with non-negative entries. The notation $\mathrm{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi)$ denotes a bivariate Poisson distribution whose marginal Poisson distributions have means $\lambda_{t,1}$ and $\lambda_{t,2}$, respectively, and φ is used for modelling the dependence between the two time series – either through a common Poisson term or a copula (see sections C.3 and C.4). It follows that $\{\boldsymbol{\lambda}_t\} = \{\boldsymbol{\lambda}_t, t \geq 1\}$ is the conditional mean process. If the largest eigenvalue (spectral radius) of $\mathbf{A}$, $\rho(A)$, is strictly less than one it follows by recursion that

$$\boldsymbol{\lambda}_t \geq (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta} \quad \text{for all } t. \tag{C.2}$$

We shall see later that this model is able to capture dependence between $Y_{t,1}$ and $Y_{t,2}$ if $\varphi \neq 0$ (section C.3) or φ does not induce the independece copula (section C.4) or if the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are not both diagonal. We notice that $\{\boldsymbol{\lambda}_t\}$ constitutes a bivariate Markov chain.

In the following sections, we consider maximum likelihood estimation (MLE) of the parameters, and for the model proposed in Liu (2012) we study the asymptotic

behaviour of the estimate. Due to the common model structure, the likelihood functions of the two models are of the same form. Let $\mathbf{A} = \{\alpha_{ij}\}_{i,j=1,2}$ and $\mathbf{B} = \{\beta_{ij}\}_{i,j=1,2}$. Then, we obtain the parameter vector $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$. The parameter space is denoted by $\Theta \subseteq \mathbb{R}^{11}$, and the true value of $\boldsymbol{\theta}$ is denoted by $\boldsymbol{\theta}_0$. The likelihood function based on the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ and conditional on $\boldsymbol{\lambda}_1$ is given by

$$L(\boldsymbol{\theta} \,|\, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\lambda}_1) = f(\mathbf{Y}_1 \,|\, \boldsymbol{\lambda}_1, \boldsymbol{\theta}) \prod_{t=2}^{n} f(\mathbf{Y}_t \,|\, \mathbf{Y}_1, \ldots, \mathbf{Y}_{t-1}, \boldsymbol{\lambda}_1, \boldsymbol{\theta}) = \prod_{t=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t)$$

with $p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t)$ being the conditional probability mass function (pmf) of the respective model ((C.3) or (C.5) below). Furthermore, the log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t).$$

C.3 Modelling of dependence through trivariate reduction

In this section, we consider the model proposed in Liu (2012), where the bivariate Poisson distribution is constructed by trivariate reduction (Mardia, 1970), where two dependent Poisson varibales are constructed from three independent Poisson variables. We adopt the setup of section C.2. Conditional on $\mathcal{F}_{t-1}$, let $X_1 \sim \text{Pois}(\lambda_{t,1} - \varphi)$, $X_2 \sim \text{Pois}(\lambda_{t,2} - \varphi)$ and $X_3 \sim \text{Pois}(\varphi)$, $0 \leq \varphi \leq \min\{\lambda_{t,1}, \lambda_{t,2}\}$, be mutually independent Poisson random variables and let $Y_{t,1} = X_1 + X_3$ and $Y_{t,2} = X_2 + X_3$. The pmf of $\mathbf{Y}_t$ is then given by

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}(Y_{t,1} = m, Y_{t,2} = n \,|\, \mathcal{F}_{t-1}) \\
&\quad = \sum_{s=0}^{m \wedge n} \mathbb{P}_{\boldsymbol{\theta}}(X_1 = m - s)\, \mathbb{P}_{\boldsymbol{\theta}}(X_2 = n - s)\, \mathbb{P}_{\boldsymbol{\theta}}(X_3 = s) \qquad \text{(C.3)} \\
&\quad = e^{-(\lambda_{t,1} + \lambda_{t,2} - \varphi)} \frac{(\lambda_{t,1} - \varphi)^m}{m!} \frac{(\lambda_{t,2} - \varphi)^n}{n!} \sum_{s=0}^{m \wedge n} \binom{m}{s} \binom{n}{s} s! \left(\frac{\varphi}{(\lambda_{t,1} - \varphi)(\lambda_{t,2} - \varphi)} \right)^s
\end{aligned}$$

with $m \wedge n = \min\{m, n\}$. For this model, the correlation structure is modelled through the common random variable X_3 with $\text{Cov}(Y_{t,1}, Y_{t,2} \,|\, \mathcal{F}_{t-1}) = \varphi$. When $\rho(A) < 1$, (C.2) provides a feasible upper bound on φ, since $0 \leq \varphi \leq \min\{\lambda_{t,1}, \lambda_{t,2}\}$ for all t. It follows that this model is only able to capture dependence between the two time series if $\varphi > 0$ or if $\mathbf{A}$ and $\mathbf{B}$ are not both diagonal. Furthermore, we notice that with the above construction of the bivariate distribution there is an upper bound on the correlation. Assume that $\lambda_{t,1} \leq \lambda_{t,2}$. Then

$$\text{Cor}(Y_{t,1}, Y_{t,2} \,|\, \mathcal{F}_{t-1}) = \frac{\varphi}{\sqrt{\lambda_{t,1} \lambda_{t,2}}} \leq \sqrt{\lambda_{t,1} / \lambda_{t,2}}.$$

If $\lambda_{t,1} << \lambda_{t,2}$ then $\text{Cor}(Y_{t,1}, Y_{t,2} \,|\, \mathcal{F}_{t-1}) << 1$. This might be the case if α_1, α_2, β_{11}, β_{12}, β_{21} and β_{22} are very small while $\delta_1 << \delta_2$.

Before continuing, we introduce some necessary terminology and definitions. We let $\|\mathbf{J}\|_p = \max_{\mathbf{x}\neq\mathbf{0}}\{\|\mathbf{J}\mathbf{x}\|_p/\|\mathbf{x}\|_p : \mathbf{x} \in \mathbb{C}^n\}$ be the p-induced norm of a general matrix $\mathbf{J} \in \mathbb{C}^{m\times n}$ for $1 \leq p \leq \infty$ with $\|\mathbf{x}\|_p$ denoting the p-norm of the vector $\mathbf{x}$. We note that $\|\mathbf{J}\|_1$ is the maximal absolute column sum of $\mathbf{J}$, and $\|\mathbf{J}\|_\infty$ is the maximal absolute row sum. The following definition introduce the *e-chain*, a property that is important in the proofs of Propositions C.2 and C.9 below.

Definition C.1. *[Meyn and Tweedie, 2009] The Markov transition function P is called equicontinuous if for any continuous function f with compact support, the sequence of functions $\{P^k f : k \in \mathbb{N}\}$ is equicontinuous on compact sets. A Markov chain which possesses an equicontinuous Markov transition function is called an e-chain.*

Liu (2012) states and proves the following proposition regarding the stability properties of the bivariate Markov chain $\{\boldsymbol{\lambda}_t\}$.

Proposition C.2 (Liu, 2012). *We apply model* (C.1) *with pmf given by* (C.3) *and assume that $\boldsymbol{\delta}, \mathbf{A}$ and $\mathbf{B}$ have non-negative entries.*

(a) If $\rho(\mathbf{A}+\mathbf{B}) < 1$, there exists at least one stationary distribution of $\{\boldsymbol{\lambda}_t\}$. In addition, if $\|\mathbf{A}\|_p < 1$ for some $1 \leq p \leq \infty$, then the stationary distribution is unique.

(b) If $\|\mathbf{A}\|_p + 2^{1-(1/p)}\|\mathbf{B}\|_p < 1$ for some $1 \leq p \leq \infty$, then $\{\boldsymbol{\lambda}_t\}$ is a geometric moment contracting Markov chain with a unique stationary and ergodic distribution, denoted by π.

The proof of the proposition is based on results from Meyn and Tweedie (2009) and Wu and Shao (2004). By showing that $\{\boldsymbol{\lambda}_t\}$ is a weak Feller chain that is bounded in probability on average, it follows from Meyn and Tweedie (2009, Theorem 12.0.1) that there exists at least one stationary distribution. Furthermore, by showing that $\{\boldsymbol{\lambda}_t\}$ is an e-chain it follows from Meyn and Tweedie (2009, Theorem 18.8.4) that the distribution is unique. The proof of (b) is based on the iterated random functions approach (see e.g., Diaconis and Freedman (1999) and Wu and Shao (2004)) and follows from Wu and Shao (2004, Theorem 2) on proving the relevant regularity conditions.

We note that the Poisson assumption is only important when showing that $\{\boldsymbol{\lambda}_t\}$ constitutes an e-chain; in the remainder of the proof, the essential properties are the structure of the model given by (C.1) and the fact that the bivariate distribution is constructed by sums of independent random variables. When proving that $\{\boldsymbol{\lambda}_t\}$ is an e-chain, the crucial feature is that $\sum_{i=0}^{\infty}|p_{\lambda_1}(i)-p_{\lambda_2}(i)| \leq 2(1-e^{-|\lambda_2-\lambda_1|})$ with p_λ denoting the pmf of a Poisson distribution with parameter λ; the left-hand side is small when $|\lambda_2 - \lambda_1|$ is small. If the Poisson distribution is written in exponential family form, the bound on the right-hand side is given by $2(1-\exp\{-|\log a(\lambda_2)-\log a(\lambda_1)|\})$ where $a(\lambda)$ is the normalising constant of the exponential family form. The proposition therefore easily generalises to a setup where the bivariate distribution is constructed through sums of independent random variables with a distribution satisfying the conditions of one of the two following setups.

(I) Let

$$f_\theta(x) = a(\theta)b(x)e^{\Phi(\theta)\cdot t(x)}, \quad x \in \mathbb{N}, \quad \theta \in \mathbb{R},$$

be the pmf of a positive, discrete random variable satisfying the convolution property $(f_{\theta_1} * f_{\theta_2})(x) = f_{\theta_1+\theta_2}(x)$ and the condition that for $\varepsilon > 0$, there exists $\eta > 0$ such that $|\log a(\theta_1) - \log a(\theta_2)| \leq \varepsilon$ when $|\theta_1 - \theta_2| \leq \eta$. Furthermore, $\Phi(\theta) \cdot t(x)$ should be increasing (or decreasing) in θ for all x.

(II) Let

$$f_{\theta,\kappa}(x) = a_\theta(\kappa) b_\theta(x) e^{\Phi(\kappa)\cdot t(x)}, \quad x \in \mathbb{N}_0, \quad \theta, \kappa \in \mathbb{R},$$

be the pmf of a positive, discrete random variable satisfying the convolution property $(f_{\theta_1,\kappa} * f_{\theta_2,\kappa})(x) = f_{\theta_1+\theta_2,\kappa}(x)$ and the condition that for $\varepsilon > 0$, there exists $\eta > 0$ such that $|\log a_{\theta_1}(\kappa) - \log a_{\theta_2}(\kappa)| \leq \varepsilon$ when $|\theta_1 - \theta_2| \leq \eta$. Furthermore, $b_\theta(x)$ should be increasing (or decreasing) in θ for all x.

For setup (I), we let $X_1 \sim f_{\theta_{t,1}-\varphi}$, $X_2 \sim f_{\theta_{t,2}-\varphi}$ and $X_3 \sim f_\varphi$, and for setup (II) we let $X_1 \sim f_{\theta_{t,1}-\varphi,\kappa}$, $X_2 \sim f_{\theta_{t,2}-\varphi,\kappa}$ and $X_3 \sim f_{\varphi,\kappa}$. For both setups, the conditions on φ depend on the choice of f.

Example C.3.
From Liu (2012) it follows that the Poisson distribution fulfills the conditions in (I).

An example of a distribution which satisfies the conditions in (II) is the negative binomial distribution with pmf given by

$$f_{\theta,\kappa}(x) = (1-\kappa)^\theta \binom{x+\theta-1}{x} e^{x\log\kappa}, \quad 0 < \kappa < 1,\ 0 < \theta < \infty.$$

It is well known that the convolution condition holds in this setting, and with $a_\theta(\kappa) = (1-\kappa)^\theta$ it follows that $|\theta_1 \log(1-\kappa) - \theta_2 \log(1-\kappa)| \leq |\theta_1 - \theta_2||\log(1-\kappa)|$. Furthermore, $\theta \mapsto \binom{x+\theta-1}{x}$ is increasing, and therefore all the conditions are met. Since $\{\boldsymbol{\lambda}_t\}$ (given by (C.1)) is the conditional mean process, $\boldsymbol{\lambda}_t = \frac{\kappa}{\kappa-1}(\theta_{t,1}, \theta_{t,2})^T$. The parameter φ should satisfy $0 \leq \varphi \leq \min\{\theta_{t,1}, \theta_{t,2}\}$ for all t, and it follows from (C.2) that $\frac{1-\kappa}{\kappa}(\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta}$ yields a feasible upper bound. ◇

With $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$, we let for the rest of this section (and *Appendix: Proofs* on page 77) $\boldsymbol{\delta}(\boldsymbol{\theta}) = (\delta_1, \delta_2)^T$, $\mathbf{A}(\boldsymbol{\theta}) = \{\alpha_{ij}\}_{i,j=1,2}$, $\mathbf{B}(\boldsymbol{\theta}) = \{\beta_{ij}\}_{i,j=1,2}$ and $\varphi(\boldsymbol{\theta}) = \varphi$. Moreover, we let $l_t(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t)$. Furthermore, we let the expected values $\mathbb{E}$ and the probabilities $\mathbb{P}$ be with respect to the stationary distribution unless otherwise stated.

Liu (2012) suggests using MLE to estimate the parameters of the model and illustrate this in an application to a real data set. In the following theorem we state that the MLE is strongly consistent under the following regularity conditions:

Assumption C.1. The log-likelihood is maximised over the set $\mathcal{D}$, where $\mathcal{D}$ is a compact subset of Θ satisfying the following:

- $\boldsymbol{\theta}_0 \in \mathcal{D}$.
- $\boldsymbol{\delta}(\boldsymbol{\theta})$, $\mathbf{A}(\boldsymbol{\theta})$ and $\mathbf{B}(\boldsymbol{\theta})$ have non-negative entries and $\mathbf{B}(\boldsymbol{\theta})$ is of full rank for all $\boldsymbol{\theta} \in \mathcal{D}$.
- $\varphi(\boldsymbol{\theta}) < (\mathbf{I} - \mathbf{A}(\boldsymbol{\theta}))^{-1}\boldsymbol{\delta}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{D}$.

- There exists a $p \in [1, \infty]$ such that $\|\mathbf{A}(\boldsymbol{\theta})\|_p + 2^{1-(1/p)}\|\mathbf{B}(\boldsymbol{\theta})\|_p < 1$ for all $\boldsymbol{\theta} \in \mathcal{D}$.

Theorem C.4. *Apply model* (C.1) *with pmf given by* (C.3). *Under Assumption C.1, the maximum likelihood estimator* $\hat{\boldsymbol{\theta}}$ *is strongly consistent, i.e.,* $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ *a.s. with respect to the stationary distribution.*

The proof is analogous to the proof of Theorem 3.1 in Wang et al. (2012), when adjusting to this two-dimensional setup, and the theorem therefore follows directly from Lemmas C.5–C.8 below. The proofs of the lemmas are deferred to *Appendix: Proofs*.

Lemma C.5. *Let the assumptions be as in Theorem C.4. Then* $\mathbb{E}[\|\boldsymbol{\lambda}_t\|_p^{ps}] < \infty$ *when* $\mathbb{E}[\|\boldsymbol{\lambda}_0\|_p^{ps}] < \infty$ *for* $s \in \mathbb{N}$, *which ensures that the stationary distribution* $\boldsymbol{\mu}$ *has moments of all orders.*

Lemma C.6. *Let the assumptions be as in Theorem C.4. Then the log-likelihood is asymptotically independent of the initial value* $\tilde{\boldsymbol{\lambda}}_1$, *i.e.,* $\sup_{\boldsymbol{\theta}\in\mathcal{D}}\left|\frac{1}{n}(l(\boldsymbol{\lambda}_1) - l(\tilde{\boldsymbol{\lambda}}_1))\right| \to 0$ *a.s. with respect to the stationary distribution.*

Lemma C.7. *Let the assumptions be as in Theorem C.4. Then* $\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t)]$ *is continuous as a function of* $\boldsymbol{\theta}$.

Lemma C.8. *Let the assumptions be as in Theorem C.4. Then the model is identifiable.*

C.4 Modelling of dependence through copulas

We again consider the setup in section C.2 but opposed to the previous section where the dependence was modelled through a common random variable, we propose here, to model it through a copula. Since copulas are defined through their cdf, the bivariate Poisson distribution of this model is also defined through its cdf and is given by

$$\mathbb{P}_{\boldsymbol{\theta}}(Y_{t,1} \leq m, Y_{t,2} \leq n \,|\, \mathcal{F}_{t-1}) = C_{\varphi}(F_{\lambda_{t,1}}(m), F_{\lambda_{t,2}}(n)), \tag{C.4}$$

where C_{φ} is a copula parameterised by φ, and F_{λ} is the cdf of a Poisson distribution with mean λ. Since $Y_{t,1}$ and $Y_{t,2}$ are integer valued random variables their joint pmf is given by

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}(Y_{t,1} = m, Y_{t,2} = n \,|\, \mathcal{F}_{t-1}) &= C_{\varphi}(F_{\lambda_{t,1}}(m), F_{\lambda_{t,2}}(n)) - C_{\varphi}(F_{\lambda_{t,1}}(m-1), F_{\lambda_{t,2}}(n)) \\ &\quad - C_{\varphi}(F_{\lambda_{t,1}}(m), F_{\lambda_{t,2}}(n-1)) + C_{\varphi}(F_{\lambda_{t,1}}(m-1), F_{\lambda_{t,2}}(n-1)).\end{aligned} \tag{C.5}$$

The conditional dependence between $Y_{t,1}$ and $Y_{t,2}$ at a given time t is modelled through the copula. Due to the properties of copulas, this definition of a bivariate Poisson distribution ensures that $Y_{t,i} \,|\, \mathcal{F}_{t-1} \sim \text{Pois}(\lambda_{t,i})$, $i = 1, 2$, regardless of the choice of copula. Model (C.1) with pmf (C.4) is able to detect dependence between the two time series $\{Y_{t,1}\}$ and $\{Y_{t,2}\}$, provided that the chosen copula is not the independence copula or if the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are not both diagonal.

A main advantage of using copulas is that with the appropriate copula, the model is able to capture negative dependence.

We remark that model (C.4) can be shown to be identifiable by using the same type of arguments as in the proof of Lemma C.8, since copulas preserve marginal distributions. In addition, this model has certain stability properties, and in this section we state and prove a result which is the equivalent of Proposition C.2. In the remainder this paper, we assume that the copula C_φ satisfies the following conditions:

(C1) $C_\varphi(u,v)$ is symmetric in u and v (this is trivial for Archimedean copulas).

(C2) $u_i \mapsto C_\varphi(u_1,u_2)$ is differentiable on $(0,1)$ for $i=1,2$.

(C3) $u_i \mapsto \frac{\partial}{\partial u_j} C_\varphi(u_1,u_2)$ is continuous on $[0,1]$ and differentiable on $(0,1)$ for $i,j=1,2$.

(C4) $\forall\, \varphi \in I_\varphi\ \exists\, 0 < K_\varphi < \infty\ \forall\, i \neq j : \frac{\partial^2}{\partial u_i \partial u_j} C_\varphi(u_1,u_2) \leq K_\varphi\ \forall\, (u_1,u_2) \in (0,1)^2$.

Under these conditions we have the following result regarding stability properties of the Markov chain $\{\boldsymbol{\lambda}_t\}$.

Proposition C.9. *Assume conditions* (C1)–(C4) *and that* $\boldsymbol{\delta}$, $\mathbf{A}$ *and* $\mathbf{B}$ *have non-negative entries. Then Proposition C.2 holds for model* (C.1) *with pmf given by* (C.5)

Since the new model is of the same type as the model considered in section C.3, the result follows if one proves that $\{\boldsymbol{\lambda}_t\}$ is an e-chain. Hence, details that coincide with the proof in Liu (2012) are omitted and instead we formulate a lemma stating that $\{\boldsymbol{\lambda}_t\}$ is an e-chain.

Lemma C.10. *Let the assumptions be as in Proposition C.9. Then* $\{\boldsymbol{\lambda}_t\}$ *is an e-chain.*

Before we prove the lemma, we give a short comment on the proof of part (b) of the proposition. As mentioned above the proof makes use of the iterated random function approach and we remark that the random function needed is

$$f_{\mathbf{u}}(\boldsymbol{\lambda}) = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda} + \mathbf{B}\widetilde{F}^{-1}_{\boldsymbol{\lambda}}(\mathbf{u}),$$

where $\boldsymbol{\lambda} = (\lambda_1,\lambda_2)^T$, $\widetilde{F}^{-1}_{\boldsymbol{\lambda}}(\mathbf{u}) = (F^{-1}_{\lambda_1}(u_1), F^{-1}_{\lambda_2}(u_2))^T \in \mathbb{N}_0^2$ and $F^{-1}_x(u) = \inf\{t \geq 0 : F_x(t) \geq u\}$. The randomness is induced by $\mathbf{u}$. It follows that, for all t $\boldsymbol{\lambda}_t = f_{\mathbf{U}_t}(\boldsymbol{\lambda}_{t-1})$, where $\{\mathbf{U}_t, t \geq 1\}$ is an independent and identically distributed sequence with distribution given by the copula C_φ.

Proof (Lemma C.10). The Markov chain $\{\boldsymbol{\lambda}_t\}$ is an e-chain, if for any continuous function f with compact support defined on $[0,\infty)\times[0,\infty)$ and $\varepsilon > 0$ there exists an $\eta > 0$ such that $|P^k_{\mathbf{x}_1} f - P^k_{\mathbf{z}_1} f| < \varepsilon$ for $\|\mathbf{x}_1 - \mathbf{z}_1\| < \eta$ and all $k \geq 1$, where $\mathbf{x}_1 = (x_{11}, x_{12})^T$, $\mathbf{z}_1 = (z_{11}, z_{12})^T$ and $\|\cdot\|$ is some norm defined on $\mathbb{R}^2$. Without loss of generality, we assume that $|f| \leq 1$.

By assumption there exists a $p \in [1,\infty]$ such that $\|A\|_p < 1$. As f is a continuous function with compact support we can take ε' and η sufficiently small such that,

whenever $\|\mathbf{x}_1 - \mathbf{z}_1\|_p < \eta$, $\varepsilon' + 8\eta K_\varphi/(1 - \|\mathbf{A}\|_p) < \varepsilon$ and $|f(\mathbf{x}_1) - f(\mathbf{z}_1)| < \varepsilon'$. In the case $k = 1$ we obtain

$$\begin{aligned}
&|P_{\mathbf{x}_1} f - P_{\mathbf{z}_1} f| \\
&\quad \leq \sum_{m,n=0}^{\infty} \left| f\left(\boldsymbol{\delta} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}(m,n)^T\right) p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1) - f\left(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}(m,n)^T\right) p_{\boldsymbol{\theta}}(m,n|\mathbf{z}_1) \right| \\
&\quad \leq \sum_{m,n=0}^{\infty} p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1) \left| f\left(\boldsymbol{\delta} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}(m,n)^T\right) - f\left(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}(m,n)^T\right) \right| \\
&\qquad + \sum_{m,n=0}^{\infty} |p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1) - p_{\boldsymbol{\theta}}(m,n|\mathbf{z}_1)| \left| f\left(\boldsymbol{\delta} + \mathbf{A}\mathbf{z}_1 + \mathbf{B}(m,n)^T\right| \right.
\end{aligned}$$

where $p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1)$ is the pmf given by (C.5). We denote the first sum on the right-hand side by S_1 and the second by S_2. The term S_2 is considered first and we observe that

$$\begin{aligned}
&\sum_{m,n=0}^{\infty} |p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1) - p_{\boldsymbol{\theta}}(m,n|\mathbf{z}_1)| \\
&\quad \leq \sum_{m,n=0}^{\infty} |p_{\boldsymbol{\theta}}(m,n|\mathbf{x}_1) - p_{\boldsymbol{\theta}}(m,n|(x_{11}, z_{12}))| + \sum_{m,n=0}^{\infty} |p_{\boldsymbol{\theta}}(m,n|(x_{11}, z_{12})) - p_{\boldsymbol{\theta}}(m,n|\mathbf{z}_1)|
\end{aligned}$$

We denote the first sum on the right-hand side by $\mathrm{S}_{2\mathrm{a}}$ and the second by $\mathrm{S}_{2\mathrm{b}}$. Due to symmetry it suffices to consider $\mathrm{S}_{2\mathrm{a}}$. Applying (C.5), $\mathrm{S}_{2\mathrm{a}}$ can be rewritten as

$$\begin{aligned}
\mathrm{S}_{2\mathrm{a}} = \sum_{m,n=0}^{\infty} \Big| & C_\varphi(F_{x_{11}}(m), F_{x_{12}}(n)) - C_\varphi(F_{x_{11}}(m-1), F_{x_{12}}(n)) \\
& - C_\varphi(F_{x_{11}}(m), F_{x_{12}}(n-1)) + C_\varphi(F_{x_{11}}(m-1), F_{x_{12}}(n-1)) \\
& - C_\varphi(F_{x_{11}}(m), F_{z_{12}}(n)) + C_\varphi(F_{x_{11}}(m-1), F_{z_{12}}(n)) \\
& + C_\varphi(F_{x_{11}}(m), F_{z_{12}}(n-1)) - C_\varphi(F_{x_{11}}(m-1), F_{z_{12}}(n-1)) \Big|.
\end{aligned}$$

Let $C_\varphi^{01}(u,v) = \frac{\partial}{\partial w} C_\varphi(u,w)\big|_{w=v}$ and $C_\varphi^{11}(u,v) = \frac{\partial^2}{\partial z \partial w} C_\varphi(z,w)\big|_{(z,w)=(u,v)}$. We then note that, for $k \in \{m-1, m\}$ and $l \in \{n-1, n\}$ it follows from the Fundamental Theorem of Calculus that

$$\begin{aligned}
&C_\varphi(F_{x_{11}}(k), F_{x_{12}}(l)) - C_\varphi(F_{x_{11}}(k), F_{z_{12}}(l)) \\
&\qquad = \{F_{x_{12}}(l) - F_{z_{12}}(l)\} \int_0^1 C_\varphi^{01}\left(F_{x_{11}}(k), F_{z_{12}}(l) + u\{F_{x_{12}}(l) - F_{z_{12}}(l)\}\right) du.
\end{aligned} \tag{C.6}$$

Now, define

$$\begin{aligned}
g(u; l, x_{12}, z_{12}) &= F_{z_{12}}(l) + u\{F_{x_{12}}(l) - F_{z_{12}}(l)\}, \\
h(v; m, x_{11}) &= F_{x_{11}}(m-1) + v\{F_{x_{11}}(m) - F_{x_{11}}(m-1)\}.
\end{aligned}$$

From (C.6), the Fundamental Theorem of Calculus and assumption (C4) we obtain

$$\begin{aligned}
&\Big|\Big[C_\varphi(F_{x_{11}}(m),F_{x_{12}}(l))-C_\varphi(F_{x_{11}}(m),F_{z_{12}}(l))\Big]\\
&\qquad-\Big[C_\varphi(F_{x_{11}}(m-1),F_{x_{12}}(l))-C_\varphi(F_{x_{11}}(m-1),F_{z_{12}}(l))\Big]\Big|\\
&\quad\le |F_{x_{12}}(l)-F_{z_{12}}(l)|\\
&\qquad\times\int_0^1\Big|C_\varphi^{01}\Big(F_{x_{11}}(m),g(u;l,x_{12},z_{12})\Big)-C_\varphi^{01}\Big(F_{x_{11}}(m-1),g(u;l,x_{12},z_{12})\Big)\Big|\,du\\
&\quad\le |F_{x_{12}}(l)-F_{z_{12}}(l)|\,p_{x_{11}}(m)\int_0^1\int_0^1\Big|C_\varphi^{11}\Big(h(v;m,x_{11}),g(u;l,x_{12},z_{12})\Big)\Big|\,dv\,du\\
&\quad\le |F_{x_{12}}(l)-F_{z_{12}}(l)|\,p_{x_{11}}(m)K_\varphi,
\end{aligned}$$

for $l\in\{n-1,n\}$, with $p_{x_{11}}$ being the Poisson pmf. Using this, we find that

$$\begin{aligned}
\mathrm{S}_{2\mathrm{a}}&\le\sum_{m,n=0}^{\infty}\Big\{|F_{x_{12}}(n)-F_{z_{12}}(n)|\,p_{x_{11}}(m)K_\varphi+|F_{x_{12}}(n-1)-F_{z_{12}}(n-1)|\,p_{x_{11}}(m)K_\varphi\Big\}\\
&=2K_\varphi|x_{12}-z_{12}|.
\end{aligned}$$

Similarly, we obtain $\mathrm{S}_{2\mathrm{b}}\le 2K_\varphi|x_{11}-z_{11}|$. We note that

$$|x_{1i}-z_{1i}|\le\|\mathbf{x}_1-\mathbf{z}_1\|_1\le c_p\|\mathbf{x}_1-\mathbf{z}_1\|_p$$

with $c_p=2^{1-(1/p)}\le 2$, for $i=1,2$. Hence, for any $\mathbf{x}_1,\mathbf{z}_1$, we have

$$\sum_{m,n=0}^{\infty}|p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{x}_1)-p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{z}_1)|\le 8K_\varphi\|\mathbf{x}_1-\mathbf{z}_1\|_p.$$

As $|f|\le 1$, we have shown that $\mathrm{S}_2\le 8K_\varphi\|\mathbf{x}_1-\mathbf{z}_1\|_p$.

Let us now look at S_1. As $\|\mathbf{A}\|_p<1$ and $\|\mathbf{x}_1-\mathbf{z}_1\|_p<\eta$ it follows that

$$\|\boldsymbol{\delta}+\mathbf{A}\mathbf{x}_1+\mathbf{B}(m,n)^T-(\boldsymbol{\delta}+\mathbf{A}\mathbf{z}_1+\mathbf{B}(m,n)^T)\|_p\le\eta,$$

and therefore $\mathrm{S}_1\le\varepsilon'$. Hence

$$|P_{\mathbf{x}_1}f-P_{\mathbf{z}_1}f|\le\varepsilon'+8K_\varphi\|\mathbf{x}_1-\mathbf{z}_1\|_p. \tag{C.7}$$

For the case $k=2$, we have

$$\begin{aligned}
&|P^2_{\mathbf{x}_1}f-P^2_{\mathbf{z}_1}f|\\
&\quad=\Big|\sum_{m,n=0}^{\infty}\Big[p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{x}_1)P_{\mathbf{x}_2}f-p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{z}_1)P_{\mathbf{z}_2}f\Big]\Big|\\
&\quad\le\sum_{m,n=0}^{\infty}p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{x}_1)|P_{\mathbf{x}_2}f-P_{\mathbf{z}_2}f|+\sum_{m,n=0}^{\infty}|p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{x}_1)-p_{\boldsymbol{\theta}}(m,n\,|\,\mathbf{z}_1)|\,|P_{\mathbf{z}_2}f|,
\end{aligned}$$

where $\mathbf{x}_2=\boldsymbol{\delta}+\mathbf{A}\mathbf{x}_1+\mathbf{B}(m,n)^T$ and $\mathbf{z}_2=\boldsymbol{\delta}+\mathbf{A}\mathbf{z}_1+\mathbf{B}(m,n)^T$. Since

$$\|\mathbf{x}_2-\mathbf{z}_2\|_p=\|\mathbf{A}(\mathbf{x}_1-\mathbf{z}_1)\|_p\le\|\mathbf{A}\|_p\|\mathbf{x}_1-\mathbf{z}_1\|_p\le\eta,$$

it follows from (C.7) that

$$\begin{aligned}|P^2_{\mathbf{x}_1} f - P^2_{\mathbf{z}_1} f| &\leq \varepsilon' + 8K_\varphi \|\mathbf{x}_2 - \mathbf{z}_2\|_p + 8K_\varphi \|\mathbf{x}_1 - \mathbf{z}_1\|_p \\ &\leq \varepsilon' + 8K_\varphi \|A\|_p \|\mathbf{x}_1 - \mathbf{z}_1\|_p + 8K_\varphi \|\mathbf{x}_1 - \mathbf{z}_1\|_p.\end{aligned}$$

Hence, by induction, we have for any $k \geq 1$

$$\begin{aligned}|P^k_{\mathbf{x}_1} f - P^k_{\mathbf{z}_1} f| &\leq \varepsilon' + 8 \sum_{s=0}^{k-1} K_\varphi \|A\|_p^s \|\mathbf{x}_1 - \mathbf{z}_1\|_p \\ &\leq \varepsilon' + \frac{8\eta K_\varphi}{1 - \|\mathbf{A}\|_p} < \varepsilon,\end{aligned}$$

proving that $\{\boldsymbol{\lambda}_t\}$ is an e-chain. □

C.4.1 Examples of copulas

As mentioned in section C.1, we consider Archimedean copulas, since these only require one parameter to model the dependence structure, and for the most common ones there exists a cdf on explicit form. We restrict attention to the Frank and Clayton copulas (Nelsen, 2006) since these are able to model negative dependence and satisfy the conditions (C1)–(C4). The cdf of the Frank copula is given by

$$C_\varphi(u,v) = \begin{cases} -\frac{1}{\varphi} \log\left(1 - \frac{(1-e^{-\varphi u})(1-e^{-\varphi v})}{1-e^{-\varphi}}\right), & \varphi \neq 0, \\ uv, & \varphi = 0, \end{cases}$$

for $(u,v) \in [0,1]^2$ and $\varphi \in (-\infty, \infty)$. We notice that $C_{-\infty}(u,v) = \max(u+v-1, 0)$, the so-called Fréchet-Hoeffding lower bound for copulas, and $C_\infty(u,v) = \min(u,v)$, the Fréchet-Hoeffding upper bound for copulas. The cdf of the Clayton copula is given by

$$\begin{aligned}C_\varphi(u,v) &= [\max(u^{-\varphi} + v^{-\varphi} - 1, 0)]^{-1/\varphi} \\ &= \begin{cases} (u^{-\varphi} + v^{-\varphi} - 1)^{-1/\varphi}, & u,v,\varphi > 0 \text{ or both } \varphi < 0 \text{ and } u^{-\varphi} + v^{-\varphi} > 1, \\ uv, & \varphi = 0, \\ 0, & \text{otherwise}, \end{cases}\end{aligned}$$

for $(u,v) \in [0,1]^2$ and $\varphi \in [-1, \infty) \setminus \{0\}$. As is the case for the Frank copula, C_{-1} is the Fréchet-Hoeffding lower bound, and C_∞ is the Fréchet-Hoeffding upper bound. We notice that for both copulas $\varphi = 0$ corresponds to independence of the marginal distributions.

To visualise the correlation struture for these copulas, we have simulated observations from the two distributions for different values of φ. The results can be seen in Figures C.1 and C.2. We notice that both models are able to capture negative dependence, although for the Clayton copula this has a somewhat special form.

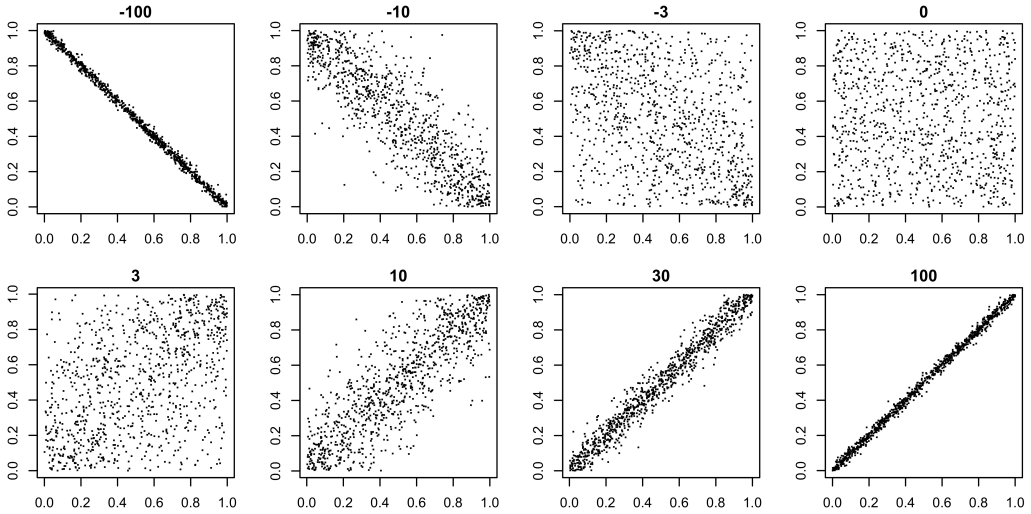


Figure C.1: Plots of simulated data from the Frank copula for different values of the parameter φ. The value of φ is indicated at the top of each plot.

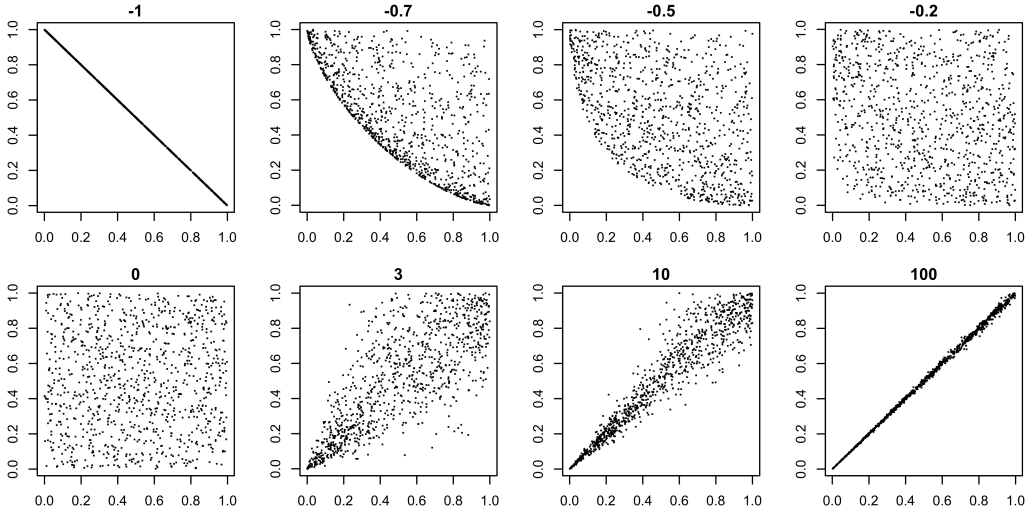


Figure C.2: Plots of simulated data from the Clayton copula for different values of the parameter φ. The value of φ is indicated at the top of each plot.

C.5 Comparison through simulation

In a simulation study, we compare the two models described in sections C.3 and C.4. The model proposed in Liu (2012) (section C.3) has a very simple structure, which makes it easy to use in practice, whereas the new model (section C.4) is slightly more advanced but in return is able to model negative dependence between the two time series if the copula is chosen appropriately. We compare four models with the structure in (C.1); two with distributions given by (C.3) and two with distributions given by (C.5). More explicitly, we compare the following models:

Model I: (C.3), $\varphi > 0$,	Model II: (C.3), $\varphi = 0$ and $\mathbf{B}$ diagonal,
Model III: (C.5), Clayton copula,	Model IV: (C.5), Frank copula.

Without loss of generality, we assume that $\mathbf{A}$ is diagonal in all four models. Then Model II corresponds to two independent time series. Under these assumptions, the Models I, III and IV all have a 9-dimensional parameter, $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$, whereas the independence Model II has a 6-dimensional parameter, $\boldsymbol{\theta} = (\delta_1, \alpha_1, \beta_1, \delta_2, \alpha_2, \beta_2)^T$.

For each model, we simulate $N = 1000$ independent replications of bivariate time series of length $n = 50, 100, 250, 500, 1000$. Data is simulated from each of the four models and for every simulated data set we check which model fit the data better. As a measure of how well the models fit the data (with respect to each other), we use the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and three different prediction scores. Most of the prediction scores presented in the literature (e.g. Czado et al. (2009) and Jung and Tremayne (2011)) are of the form $(n-1)^{-1}\sum_{t=2}^{n} s(F_t(Y_t))$, where $s(\cdot)$ is the scoring rule and $F_t(\cdot)$ is the cdf of the predicted distribution. We will consider the following prediction scores: the logarithmic score (LS), the quadratic score (QS) and the ranked probability score (RPS). To reduce the computational costs, one-dimensional scores have been used. The scores are defined as follows:

$$\begin{aligned}
\text{LS}: &\quad s(F_t(Y_t)) = -\log p_t(Y_t), \\
\text{QS}: &\quad s(F_t(Y_t)) = -2p_t(Y_t) + \|p_t\|^2, \\
\text{RPS}: &\quad s(F_t(Y_t)) = \sum_{j=0}^{\infty} \{F_t(j) - \mathbf{1}_{\{Y_t \le j\}}\}^2,
\end{aligned}$$

where $p_t(\cdot)$ is the pmf of the predicted distribution and $\|p_t\|^2 = \sum_{j=0}^{\infty} p_t(j)^2$. In general one seeks to minimise these measures.

The data are generated using the algorithms described below. When simulating a bivariate time series of length n, distributed according to the model (C.1), 500 extra observations are simulated in order to simulate from the stationary distribution (Algorithm 1).

Algorithm 1. Generation of a bivariate time series of length n distributed according to the model (C.1), where $\text{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi)$ refers to the distribution corresponding to either (C.3) or (C.4).

1. Set $\boldsymbol{\lambda}_1 = (1,1)^T$ and simulate $\mathbf{X}_1 = (X_{1,1}, X_{1,2})^T$ from $\text{BP}(1,1,\varphi)$.

2. Repeat for $t = 2, \ldots, n + 500$:
   (a) Set $\boldsymbol{\lambda}_t = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{X}_{t-1}$.
   (b) Simulate $\mathbf{X}_t = (X_{t,1}, X_{t,2})^T$ from $\text{BP}(\lambda_{t,1}, \lambda_{t,2}, \varphi)$.
3. Set $\mathbf{Y} = (\mathbf{X}_{501} \cdots \mathbf{X}_{n+500})$.

The following two algorithms specify how the simulation in 2.(b) of Algorithm 1 can be carried out, depending on the desired pmf.

Algorithm 2. Generation of a random vector $\mathbf{Y}$ from the distribution $\text{BP}(\lambda_1, \lambda_2, \varphi)$ corresponding to (C.3):

1. Simulate $X_1 \sim \text{Pois}(\lambda_1 - \varphi)$, $X_2 \sim \text{Pois}(\lambda_2 - \varphi)$ and $X_3 \sim \text{Pois}(\varphi)$.
2. Set $\mathbf{Y} = (X_1 + X_3, X_2 + X_3)^T$.

Algorithm 3. Generation of a random vector $\mathbf{Y}$ from the distribution $\text{BP}(\lambda_1, \lambda_2, \varphi)$ corresponding to (C.5):

1. Simulate (U, V) from the copula distribution C_φ. Suppose $(U, V) = (u, v)$.
2. Calculate $y_1 = F^{-1}_{\lambda_1}(u)$ and $y_2 = F^{-1}_{\lambda_2}(v)$.
3. Set $\mathbf{Y} = (y_1, y_2)^T$.

C.5.1 Results of the simulations

The outcomes of the simulations are presented in Tables C.3–C.10 in *Appendix: Results of the simulations* on pages 82–85. Tables C.3, C.5, C.7 and C.9 report the empirical bias, the standard deviation (std) of the bias and the root mean squared error (rmse) for the parameters of the model being simulated from. When simulating from Model II, which, as already mentioned, corresponds to simulating two independent time series, one would expect that the three other models estimate the dependence parameter φ to be close to zero. When $\varphi = 0$, the Frank and Clayton copulas both correspond to the independence copula. Therefore, the bias, std of the bias and the (rmse) have also been reported for the estimate of φ for these models (Table C.5). To visualise the results, the bias and the std have been plotted as functions of n in Figures C.3–C.6. From these figures it is clear that for all four models the bias decreases with increasing sample size and that the bias is small compared to the standard deviation.

As mentioned above, five different measures, AIC, BIC, LS, QS and RPS, are used in order to determine how well the model of interest fits the simulated data. To make use of these measures, we have reported the fraction of the 1000 runs where the given model provides the best fit to the data or provides the worst fit to the data, respectively. The results are presented in Tables C.4, C.6, C.8 and C.10. A visualisation of the results can be seen in Figure C.7 on page 75. The figure shows plots of the fraction of 'best fit' versus the sample size n for each measure for each fitted model. Each column corresponds to simulations from one

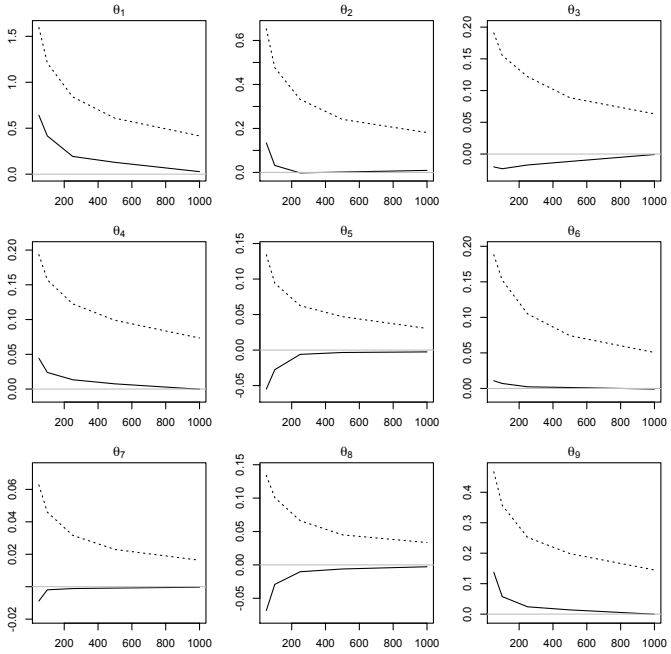


Figure C.3: Illustration of the results of the simulation from Model I reported in Table C.3, where full lines represent the bias and dashed ones the std. Recall that $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$.

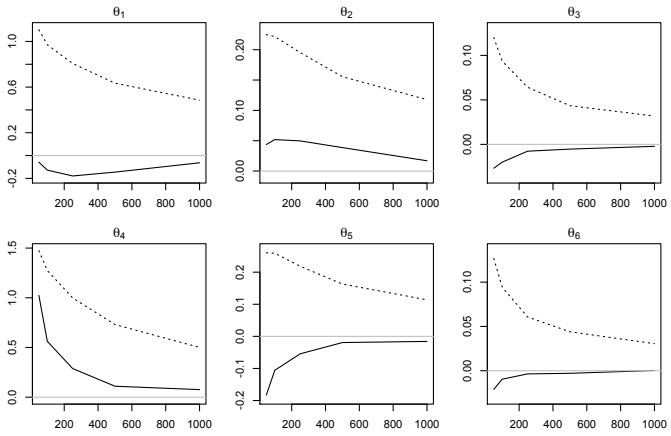


Figure C.4: Illustration of the results of the simulation from Model II reported in Table C.5, where full lines represent the bias and dashed ones the std. Recall that $\boldsymbol{\theta} = (\delta_1, \alpha_1, \beta_1, \delta_2, \alpha_2, \beta_2)^T$.

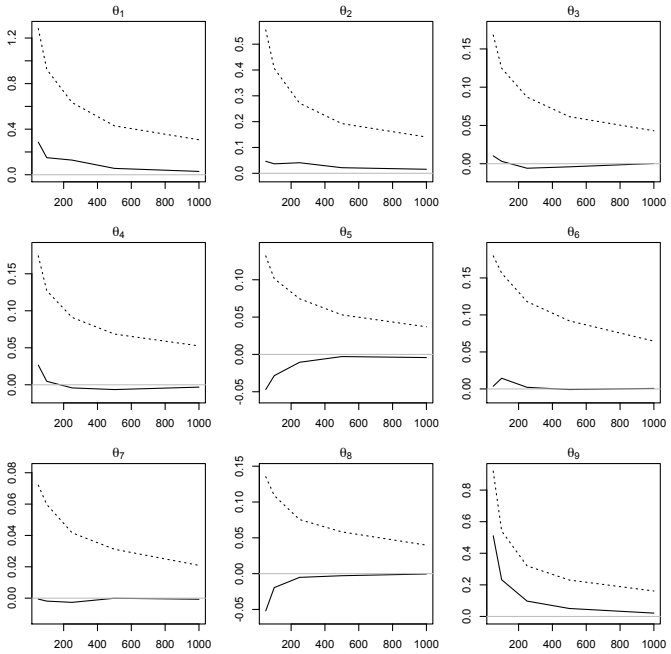


Figure C.5: Illustration of the results of the simulation from Model III reported in Table C.7, where full lines represent the bias and dashed ones the std. Recall that $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$.

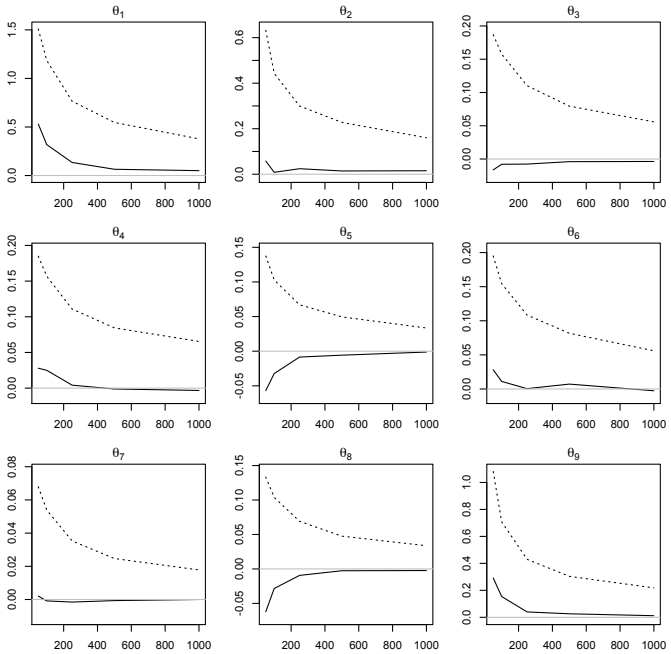


Figure C.6: Illustration of the results of the simulation from Model IV reported in Table C.9, where full lines represent the bias and dashed ones the std. Recall that $\boldsymbol{\theta} = (\delta_1, \delta_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \varphi)^T$.

model, with the leftmost from Model I and the rightmost from Model IV. Each row represents one measure, with the top row corresponding to the AIC and the bottom row corresponding to the RPS. The colour represents the model that is being fitted.

The results from the copula-based models are shown in the two rightmost columns. It is clear that for both models the true model has the highest fraction of best fit for all sample sizes and the fraction increases with increasing sample size. The trend is more pronounced for Model III where the fraction of best fit reaches 1000 out of 1000 when $n = 1000$. The results from Model I are shown in the leftmost column. It follows that, based on the scores, Model I, III and IV fit the data equally well for all sample sizes, whereas the likelihood-based measures favour the true model when the sample size is large but not for small sample sizes. For Model II, the AIC and the BIC both favour the true model whereas the scores are difficult to interpret. To summarise, these results indicate that for a small sample size the copula-based models (in particular the Clayton) provide a better fit.

C.6 Data application

We consider a data set containing the number of daytime (6:00am - 10:00pm) and nighttime (10:00pm - 6:00am) road accidents in the Schipol area in the Netherlands (Pedeli and Karlis, 2011). The data are shown in Figure C.8. In this section, we fit the four models considered in the simulation study to the data and use the same five measures to decide which one provides the best fit to the data. The results of the fits can be seen in Table C.1.

As previously mentioned, each of the measures favours the model with the lowest value, and we see from Table C.1 that all five measures favour Model I. We also notice that for all measures the values for all models seem to be very close. To investigate if the values are in fact close, corresponding to the models fitting equally well, we simulated 1000 observations under Model I with the maximum likelihood estimate under Model I, $\hat{\theta}_I = (2.06, 0.70, 0.56, 0.28, 0.08, 0.37, 0.04, 0.06, 0.26)^T$, as parameter. The results are reported in Table C.2. We see that the likelihood measures favour Model I and the score measures favour Model III whereas Model II has the highest fraction of worst fit with respect to all five measures. This could indicate that both Model I and Model III fit the data equally well and that the two time series are indeed dependent.

Table C.1: Results of the model fit.

Model	AIC	BIC	LS	QS	RPS
I	3472.911	3508.010	2.3691	−0.1449	1.6323
II	3496.228	3525.860	2.3888	−0.1433	1.6529
III	3484.689	3519.789	2.3777	−0.1443	1.6379
IV	3496.228	3531.327	2.3828	−0.1442	1.6491

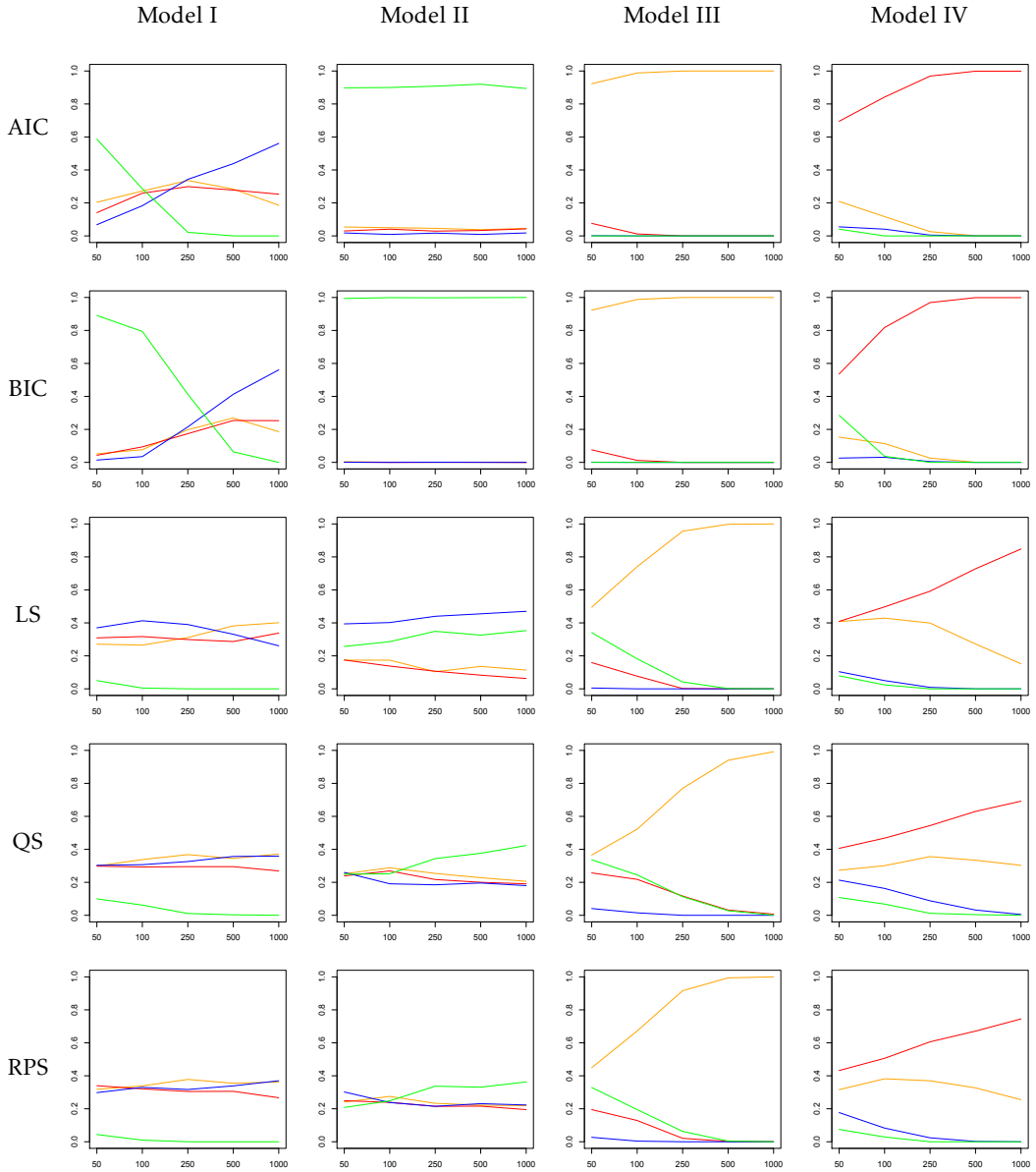


Figure C.7: Visualization of the simulation results. Each plot shows 'best fit' as a function of n. Each column corresponds to simulations from one model with the leftmost from Model I and the rightmost from Model IV. Each row represents one measure with the AIC at the top and the RPS at the bottom. The colour of the graphs refer to the model being fitted to the data (blue: I, green: II, orange: III, red: IV).

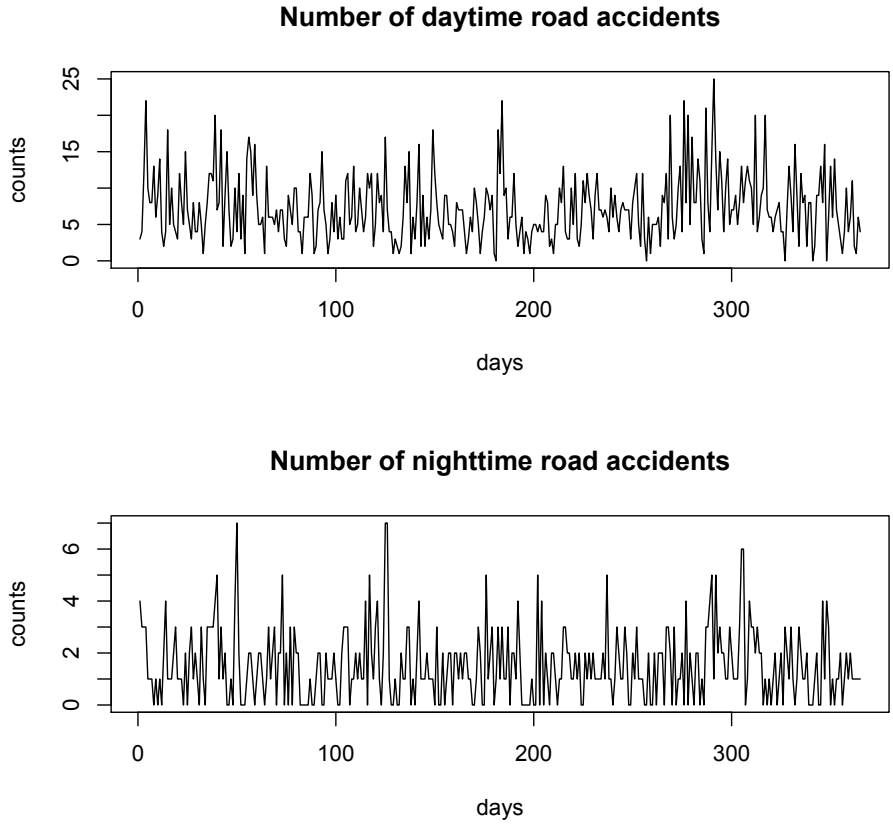


Figure C.8: Number of daytime and nighttime road accidents in the Schiphol area in the Netherlands.

Table C.2: Results of the 1000 simulations from Model I with sample size $n = 365$ and parameter $\boldsymbol{\theta} = (2.06, 0.70, 0.56, 0.28, 0.08, 0.37, 0.04, 0.06, 0.26)^T$. For each model, an entry in the table reports the fraction of the 1000 runs for which the given measure has selected the specific model as the best or the worst fit, respectively.

	Best fit					Worst fit				
Model	AIC	BIC	LS	QS	RPS	AIC	BIC	LS	QS	RPS
I	**0.422**	**0.319**	0.149	0.266	0.197	0.003	0.099	0.000	0.003	0.000
II	0.007	0.278	0.000	0.003	0.000	**0.982**	**0.644**	**0.995**	**0.989**	**0.995**
III	0.281	0.207	**0.478**	**0.386**	**0.414**	0.006	0.194	0.000	0.001	0.000
IV	0.290	0.196	0.373	0.345	0.389	0.009	0.063	0.005	0.007	0.005

C.7 Future work

This paper is a work in progress. Asymptotic results for the copula-based model remains to be investigated as well as asymptotic normality for the model of section C.3. When considering the copula-based model, the form of the pmf, given in (C.5), is a challenge. Furthermore, application of the two types of models to several real data examples, and further simulation studies, would be desirable in order to explore the advantages and disadvantages of the two models in details. The preliminary results obtained in this paper, seem to indicate that the copula-based model gives a better fit for small sample sizes.

Acknowledgements

The authors wish to thank Heng Liu and Johan Segers for fruitful discussions.

Appendix: Proofs

Proof (Lemma C.5). Let $\boldsymbol{\delta} = (\delta_1, \delta_2)^T$, $\mathbf{A} = \{\alpha_{ij}\}_{i,j=1,2}$ and $\mathbf{B} = \{\beta_{ij}\}_{i,j=1,2}$. Furthermore, let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T \in \mathbb{R}_+^2$, with $\boldsymbol{\gamma} \geq (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta}$. Now, let $s \geq 1$ and $V(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|_p^{ps}$. We note that $\lambda_{1,i} = \delta_i + \alpha_{i1}\lambda_{0,1} + \alpha_{i2}\lambda_{0,2} + \beta_{i1}Y_{0,1} + \beta_{i2}Y_{0,2}$ for $i = 1, 2$. Then

$$\begin{aligned}
\frac{\mathbb{E}[V(\boldsymbol{\lambda}_1)\,|\,\boldsymbol{\lambda}_0 = \boldsymbol{\gamma}]}{V(\boldsymbol{\gamma})} &= \mathbb{E}\left[\left\{\frac{\lambda_{1,1}^p + \lambda_{1,2}^p}{\|\boldsymbol{\gamma}\|_p^p}\right\}^s \,\middle|\, \boldsymbol{\lambda}_0 = \boldsymbol{\gamma}\right] \\
&= \mathbb{E}\left[\left\{\left(\frac{\delta_1 + \alpha_{11}\gamma_1 + \alpha_{12}\gamma_2 + \beta_{11}Y_{0,1} + \beta_{12}Y_{0,2}}{\|\boldsymbol{\gamma}\|_p}\right)^p \right.\right. \\
&\qquad \left.\left. + \left(\frac{\delta_2 + \alpha_{21}\gamma_1 + \alpha_{22}\gamma_2 + \beta_{21}Y_{0,1} + \beta_{22}Y_{0,2}}{\|\boldsymbol{\gamma}\|_p}\right)^p\right\}^s\right] \\
&= \mathbb{E}[h(\mathbf{Y}_0, \boldsymbol{\gamma})].
\end{aligned}$$

Let $u_i = \gamma_i/\|\boldsymbol{\gamma}\|_p$, $i = 1, 2$, and $\boldsymbol{u} = (u_1, u_2)^T$. We notice that $Y_{0,i} \sim \text{Po}(\gamma_i)$, $i = 1, 2$, conditionally on $\{\boldsymbol{\lambda}_0 = \boldsymbol{\gamma}\}$. If $Z \sim \text{Po}(\lambda)$, we have for $k \geq 1$ $\mathbb{E}[|Z - \lambda|^k]^{1/k} \leq O(1 + \lambda^{1/2})$. This gives

$$\mathbb{E}\left[\left|u_i \frac{Y_{0,i} - \gamma_{i,0}}{\gamma_i}\right|^k\right]^{1/k} \leq O\left(u_i \frac{1 + \gamma_i^{1/2}}{\gamma_i}\right) \to 0,$$

for $\|\boldsymbol{\gamma}\|_p \to \infty$. From this we find that

$$\mathbb{E}[\,|h(\mathbf{Y}_0, \boldsymbol{\gamma}) - h(\boldsymbol{\gamma}, \boldsymbol{\gamma})|\,] \to 0 \quad \text{for} \quad \|\boldsymbol{\gamma}\|_p \to \infty.$$

Furthermore, it follows from the assumptions that

$$h(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = \{\|(\mathbf{A} + \mathbf{B})\boldsymbol{u}\|_p^p\}^s \leq \{\|\mathbf{A}\|_p + 2^{1-(1/p)}\|\mathbf{B}\|_p\}^{ps} < 1.$$

The result now follows from the same type of arguments as in Wang et al. (2012). See also Proposition 6.2.12 and the remarks in section 6.2.2 in Duflo (1997). □

Proof (Lemma C.6). We notice that

$$\boldsymbol{\lambda}_t = \boldsymbol{\delta} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}Y_{t-1} = \mathbf{C}_{t-1} + \mathbf{A}\boldsymbol{\lambda}_{t-1} = \sum_{k=1}^{t-1} \mathbf{A}^{k-1}\mathbf{C}_{t-k} + \mathbf{A}^{t-1}\boldsymbol{\lambda}_1$$

with $\mathbf{C}_t = \boldsymbol{\delta} + \mathbf{B}Y_t$ which implies that $\boldsymbol{\lambda}_t$ can be expressed as a function of $\boldsymbol{\lambda}_1$. Let $\rho = \sup_{\boldsymbol{\theta}\in\mathcal{D}} \|\mathbf{A}\|_p < 1$. Then

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\mathcal{D}} \|\boldsymbol{\lambda}_t(\boldsymbol{\lambda}_1) - \tilde{\boldsymbol{\lambda}}_t(\tilde{\boldsymbol{\lambda}}_1)\|_p \\
&\quad = \sup_{\boldsymbol{\theta}\in\mathcal{D}} \left\| \sum_{k-1}^{t-1} \mathbf{A}^{k-1}\mathbf{C}_{t-k} + \mathbf{A}^{t-1}\boldsymbol{\lambda}_1 - \sum_{k-1}^{t-1} \mathbf{A}^{k-1}\mathbf{C}_{t-k} + \mathbf{A}^{t-1}\tilde{\boldsymbol{\lambda}}_1 \right\|_p \\
&\quad = \sup_{\boldsymbol{\theta}\in\mathcal{D}} \|\mathbf{A}^{t-1}(\boldsymbol{\lambda}_1 - \tilde{\boldsymbol{\lambda}}_1)\|_p \leq \rho^{t-1}\|\boldsymbol{\lambda}_1 - \tilde{\boldsymbol{\lambda}}_1\|_p = K\rho^t,
\end{aligned}$$

with $K = \|\boldsymbol{\lambda}_1 - \tilde{\boldsymbol{\lambda}}_1\|_p/\rho$. From the form of the log-likelihood function it follows that

$$\sup_{\boldsymbol{\theta}\in\mathcal{D}} |\tfrac{1}{n}(l(\boldsymbol{\lambda}_1) - l(\tilde{\boldsymbol{\lambda}}_1))| \leq \sup_{\boldsymbol{\theta}\in\mathcal{D}} \tfrac{1}{n} \sum_{t=1}^{n} |\log p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_t) - \log p_{\boldsymbol{\theta}}(\mathbf{Y}_t \,|\, \tilde{\boldsymbol{\lambda}}_t)|. \tag{C.8}$$

To ease the notation, we let $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$, $\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}_t$ and $\mathbf{Y} = \mathbf{Y}_t$. Let

$$f(\boldsymbol{\lambda}, \varphi) = \frac{\varphi}{(\lambda_1 - \varphi)(\lambda_2 - \varphi)} \quad \text{and} \quad g(\mathbf{Y}, \boldsymbol{\lambda}, \varphi) = \sum_{s=0}^{Y_1 \wedge Y_2} \binom{Y_1}{s}\binom{Y_2}{s} s! f(\boldsymbol{\lambda}, \varphi)^s,$$

and assume that $f(\tilde{\boldsymbol{\lambda}}, \varphi) \geq f(\boldsymbol{\lambda}, \varphi)$ (when $f(\tilde{\boldsymbol{\lambda}}, \varphi) < f(\boldsymbol{\lambda}, \varphi)$ a similar argument can be made). For a term in (C.8), we reach the upper bound

$$\begin{aligned}
&|\log p_{\boldsymbol{\theta}}(\mathbf{Y} \,|\, \boldsymbol{\lambda}) - \log p_{\boldsymbol{\theta}}(\mathbf{Y} \,|\, \tilde{\boldsymbol{\lambda}})| \\
&\quad = \Big| -(\lambda_1 + \lambda_2 - \varphi) + Y_1 \log(\lambda_1 - \varphi) + Y_2 \log(\lambda_2 - \varphi) + \log g(\mathbf{Y}, \boldsymbol{\lambda}, \varphi) \\
&\qquad + (\tilde{\lambda}_1 + \tilde{\lambda}_2 - \varphi) - Y_1 \log(\tilde{\lambda}_1 - \varphi) - Y_2 \log(\tilde{\lambda}_2 - \varphi) - \log g(\mathbf{Y}, \tilde{\boldsymbol{\lambda}}, \varphi) \Big| \\
&\quad \leq 2K\rho^t + Y_1 \left| \log\left(\frac{\tilde{\lambda}_1 - \varphi}{\lambda_1 - \varphi} \right) \right| + Y_2 \left| \log\left(\frac{\tilde{\lambda}_2 - \varphi}{\lambda_2 - \varphi} \right) \right| + \left| \log\left(\frac{g(\mathbf{Y}, \tilde{\boldsymbol{\lambda}}, \varphi)}{g(\mathbf{Y}, \boldsymbol{\lambda}, \varphi)} \right) \right|.
\end{aligned} \tag{C.9}$$

We look at the last three terms in this sum in turn. For the first two terms, we obtain the upper bound

$$Y_i \left| \log\left(\frac{\tilde{\lambda}_i - \varphi}{\lambda_i - \varphi} \right) \right| \leq Y_i \frac{|\lambda_i - \varphi - (\tilde{\lambda}_i - \varphi)|}{\varepsilon} \leq \|\mathbf{Y}\|_2 \frac{K\rho^t}{\varepsilon}, \quad i = 1, 2, \tag{C.10}$$

with $\varepsilon = \inf_{\boldsymbol{\theta}\in\mathcal{D}}(\min_j((\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\delta})_j - \varphi)$. Applying the same type of inequalities, an

upper bound for the last term is found to be

$$
\begin{aligned}
&\left|\log\left(\frac{g(\mathbf{Y},\tilde{\boldsymbol{\lambda}},\varphi)}{g(\mathbf{Y},\boldsymbol{\lambda},\varphi)}\right)\right| \\
&\quad= \log\left(\frac{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!\exp\{s\log f(\tilde{\boldsymbol{\lambda}},\varphi)\}}{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!f(\boldsymbol{\lambda},\varphi)^s}\right) \\
&\quad= \log\left(\frac{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!f(\boldsymbol{\lambda},\varphi)^s\exp\{s[\log f(\tilde{\boldsymbol{\lambda}},\varphi)-\log f(\boldsymbol{\lambda},\varphi)]\}}{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!f(\boldsymbol{\lambda},\varphi)^s}\right) \\
&\quad\le \log\left(\exp\Big\{(Y_1\wedge Y_2)[\log f(\tilde{\boldsymbol{\lambda}},\varphi)-\log f(\boldsymbol{\lambda},\varphi)]\Big\}\frac{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!f(\boldsymbol{\lambda},\varphi)^s}{\sum_{s=0}^{Y_1\wedge Y_2}\binom{Y_1}{s}\binom{Y_2}{s}s!f(\boldsymbol{\lambda},\varphi)^s}\right) \\
&\quad\le (Y_1\wedge Y_2)\log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\varphi)}{f(\boldsymbol{\lambda},\varphi)}\right) \\
&\quad\le 2\|\mathbf{Y}\|_2\frac{K\rho^t}{\varepsilon}. \qquad \text{(C.11)}
\end{aligned}
$$

It then follows that the difference between the log-likelihoods based on an arbitrary initial value and on the stationary initial one is

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}\in\mathcal{D}}|\tfrac{1}{n}(l(\boldsymbol{\lambda}_1)-l(\tilde{\boldsymbol{\lambda}}_1))| &\le \sup_{\boldsymbol{\theta}\in\mathcal{D}}\tfrac{1}{n}\sum_{t=1}^{n}\left\{2K\rho^t+2\|\mathbf{Y}_t\|_2\frac{K\rho^t}{\varepsilon}+2\|\mathbf{Y}_t\|_2\frac{K\rho^t}{\varepsilon}\right\} \\
&\le \tfrac{1}{n}c(K,\varepsilon)\sum_{t=1}^{n}\rho^t(1+\|\mathbf{Y}_t\|_2),
\end{aligned}
$$

with $c(K,\varepsilon)$ a constant depending on K and ε but independent of $\mathbf{Y}_t$. The result now follows from Cesàro's Lemma if $\rho^t\|\mathbf{Y}_t\|_2\to 0$ a.s. The Borel-Cantelli Lemma gives that it suffices to show that for all $\varepsilon>0$ we have $\sum_{t=1}^{\infty}\mathbb{P}(\rho^t\|Y_t\|_2>\varepsilon)<\infty$. From Markov's inequality with $s\in\mathbb{N}$ we obtain that

$$
\sum_{t=1}^{\infty}\mathbb{P}(\rho^t\|\mathbf{Y}_t\|_2>\varepsilon)\le\sum_{t=1}^{\infty}\frac{\mathbb{E}[(\rho^t\|\mathbf{Y}_t\|_2)^s]}{\varepsilon^s}\le\tfrac{1}{\varepsilon^s}\sum_{t=1}^{\infty}\rho^{st}\,\mathbb{E}[\mathrm{poly}(\boldsymbol{\lambda}_t,s)],
$$

with $\mathrm{poly}(\boldsymbol{\lambda}_t,s)$ being a polynomium in $\lambda_{t,1}$ and $\lambda_{t,2}$ of order s. It follows from Lemma C.5 that this sum is finite (see also Francq and Zakoian (2004)). □

Proof (Lemma C.7). For $\boldsymbol{\theta}\in\mathcal{D}$ let $V_\eta(\boldsymbol{\theta})\subset\mathcal{D}$ be an open ball with radius η centered at $\boldsymbol{\theta}$. The result follows if

$$
\mathbb{E}\Big[\sup_{\tilde{\boldsymbol{\theta}}\in V_\eta(\boldsymbol{\theta})}|l_t(\tilde{\boldsymbol{\theta}})-l_t(\boldsymbol{\theta})|\Big]\to 0\quad\text{for}\quad\eta\to 0.
$$

We adopt the notation from the proof of Lemma C.6 and let $c(\cdot)$ be a generic, positive, finite constant depending on its arguments. Then, by using inequalities

for the p-norm

$$
\begin{aligned}
&\|\boldsymbol{\lambda}_t(\boldsymbol{\theta}) - \boldsymbol{\lambda}_t(\tilde{\boldsymbol{\theta}})\|_p \\
&\quad = \Big\| \sum_{k=1}^{t-1} \mathbf{A}^{k-1}\mathbf{C}_{t-k} + \mathbf{A}^{t-1}\boldsymbol{\lambda}_1 - \sum_{k=1}^{t-1} \tilde{\mathbf{A}}^{k-1}\tilde{\mathbf{C}}_{t-k} - \tilde{\mathbf{A}}^{t-1}\boldsymbol{\lambda}_1 \Big\|_p \\
&\quad \leq \sum_{k=1}^{t-1} \Big\{ \|(\mathbf{A}^{k-1} - \tilde{\mathbf{A}}^{k-1})\mathbf{C}_{t-k}\|_p + \|\tilde{\mathbf{A}}^{k-1}(\mathbf{C}_{t-k} - \tilde{\mathbf{C}}_{t-k})\|_p \Big\} + \|(\mathbf{A}^{t-1} - \tilde{\mathbf{A}}^{t-1})\boldsymbol{\lambda}_1\|_p \\
&\quad \leq \sum_{k=1}^{t-1} \Big\{ 2(k-1)\rho^{k-2}\eta \|\mathbf{C}_{t-k}\|_p + \rho^{k-1}\Big[\|\delta - \tilde{\delta}\|_p + \|(\mathbf{B} - \tilde{\mathbf{B}})\mathbf{Y}_{t-k}\|_p \Big] \Big\} \\
&\qquad + 2(t-1)\rho^{t-2}\eta \|\boldsymbol{\lambda}_1\|_p \\
&\quad \leq \sum_{k=1}^{t-1} \Big\{ 2(k-1)\rho^{k-2}\eta (\|\delta\|_p + \|\mathbf{B}\mathbf{Y}_{t-k}\|_p) + \rho^{k-1}\Big[2\eta + 2\eta \|\mathbf{Y}_{t-k}\|_p \Big] \Big\} \\
&\qquad + 2(t-1)\rho^{t-2}\eta \|\boldsymbol{\lambda}_1\|_p \\
&\quad \leq \eta c(\delta, \mathbf{B}, \rho) \sum_{k=1}^{t-1} \Big\{ \|\boldsymbol{\lambda}_1\|_p + k\rho^{k-1}(1 + \|\mathbf{Y}_{t-k}\|_p) \Big\},
\end{aligned}
$$

with $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$, $\tilde{\mathbf{A}} = \mathbf{A}(\tilde{\boldsymbol{\theta}})$, etc. We notice that the right hand side is independent of $\tilde{\boldsymbol{\theta}}$.

For ease of notation, let $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_t(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\lambda}}_t = \boldsymbol{\lambda}_t(\tilde{\boldsymbol{\theta}})$. Assuming that $f(\tilde{\boldsymbol{\lambda}}, \tilde{\varphi}) \geq f(\boldsymbol{\lambda}, \varphi)$ and following the same method as in the proof of Lemma C.6 (see (C.9)–(C.11)) we find that

$$
\begin{aligned}
&|l_t(\tilde{\boldsymbol{\theta}}) - l_t(\boldsymbol{\theta})| \\
&\quad \leq \|\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t\|_p + \eta + 2\|\mathbf{Y}_t\|_p \frac{\|\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t\|_p + \eta}{\varepsilon} + \|\mathbf{Y}_t\|_p \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_t, \tilde{\varphi})}{f(\boldsymbol{\lambda}_t, \varphi)} \right) \\
&\quad \leq c(\varepsilon)\{1 + \|\mathbf{Y}_t\|_p\}\{\|\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t\|_p + \eta\} \\
&\quad \leq c(\varepsilon)\{1 + \|\mathbf{Y}_t\|_p\}\Big\{ \eta + \eta c(\delta, \mathbf{B}, \rho) \sum_{k=1}^{t-1} \Big\{ \|\boldsymbol{\lambda}_1\|_p + k\rho^{k-1}(1 + \|\mathbf{Y}_{t-k}\|_p) \Big\} \Big\} \\
&\quad \leq \eta c(\delta, \mathbf{B}, \varepsilon, \rho)\{1 + \|\mathbf{Y}_t\|_p\}\Big\{ 1 + \sum_{k=1}^{t-1} \Big\{ \|\boldsymbol{\lambda}_1\|_p + k\rho^{k-1}(1 + \|\mathbf{Y}_{t-k}\|_p) \Big\} \Big\}.
\end{aligned}
$$

Again, we notice that the right-hand side is independent of $\tilde{\boldsymbol{\theta}}$, and a similar argument can be made if $f(\tilde{\boldsymbol{\lambda}}, \tilde{\varphi}) < f(\boldsymbol{\lambda}, \varphi)$. It follows from Hölder's inequality that

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{\tilde{\boldsymbol{\theta}} \in V_\eta(\boldsymbol{\theta})} |l_t(\tilde{\boldsymbol{\theta}}) - l_t(\boldsymbol{\theta})| \Big] \\
&\quad \leq \eta c(\delta, \mathbf{B}, \varepsilon, \rho)\, \mathbb{E}\Big[|1 + \|\mathbf{Y}_t\|_p|^p \Big]^{1/p} \mathbb{E}\Big[\Big| 1 + \sum_{k=1}^{t-1} \Big\{ \|\boldsymbol{\lambda}_1\|_p + k\rho^{k-1}(1 + \|\mathbf{Y}_{t-k}\|_p) \Big\} \Big|^p \Big]^{1/p}.
\end{aligned}
$$

The two expected values are finite and the result follows from applying Lemma C.5 and Hölder's and Minkowski's inequalities. □

Proof (Lemma C.8). Let $\mathbb{E}_{\theta_0}$ be the expected value with respect to the true value of θ. From Jensen's inequality, it follows that

$$\begin{aligned}\mathbb{E}_{\theta_0}[l_t(\theta) - l_t(\theta_0)] &= \mathbb{E}_{\theta_0}\Big[\mathbb{E}_{\theta_0}\Big(\log\Big\{\frac{p_{\theta_0}(\mathbf{Y}_t \,|\, \lambda_t(\theta))}{p_{\theta_0}(\mathbf{Y}_t \,|\, \lambda_t(\theta_0))}\Big\}\,\Big|\,\mathcal{F}_{t-1}\Big)\Big] \\ &\leq \mathbb{E}_{\theta_0}\Big[\log \mathbb{E}_{\theta_0}\Big(\frac{p_{\theta}(\mathbf{Y}_t \,|\, \lambda_t(\theta))}{p_{\theta_0}(\mathbf{Y}_t \,|\, \lambda_t(\theta_0))}\,\Big|\,\mathcal{F}_{t-1}\Big)\Big] \\ &= \mathbb{E}_{\theta_0}(\log 1) \\ &= 0,\end{aligned}$$

with equality if and only if $p_{\theta}(\mathbf{Y}_t \,|\, \lambda_t(\theta)) = p_{\theta_0}(\mathbf{Y}_t \,|\, \lambda_t(\theta_0))$ a.s. $\mathcal{F}_{t-1}$. Thus $\mathbb{E}_{\theta_0}[l(\theta)] \leq \mathbb{E}_{\theta_0}[l(\theta_0)]$, and equality implies $\mathbb{E}_{\theta_0}[l_t(\theta)] = \mathbb{E}_{\theta_0}[l_t(\theta_0)]$, $t = 1,\dots,n$, that is $\varphi = \varphi_0$ and $\lambda_t(\theta) = \lambda_t(\theta_0)$ a.s. $\mathcal{F}_{t-1}$, $t = 1,\dots,n$. Assume that $\tilde{\theta}$ satisfies $\lambda_t(\tilde{\theta}) = \lambda_t(\theta_0)$ a.s. $\mathcal{F}_{t-1}$. This gives

$$0 = \lambda_t(\tilde{\theta}) - \lambda_t(\theta_0) = (\tilde{\delta}(\tilde{\theta}) - \delta(\theta_0)) + (\mathbf{A}(\tilde{\theta}) - \mathbf{A}(\theta_0))\lambda_{t-1} + (\mathbf{B}(\tilde{\theta}) - \mathbf{B}(\theta_0))\mathbf{Y}_{t-1},$$

and varying $\mathbf{Y}_{t-1}$ shows that $\mathbf{B}(\tilde{\theta}) - \mathbf{B}(\theta_0) = 0$. Also, when $\mathbf{B}(\theta_0)$ has full rank, λ_{t-1} is not confined to a lower-dimensional subspace of $\mathbb{R}^2$, and so $\mathbf{A}(\tilde{\theta}) = \mathbf{A}(\theta_0)$ and $\tilde{\delta}(\tilde{\theta}) - \delta(\theta_0)$. □

Appendix: Results of the simulations

The Tables C.3–C.10 on pages 82–85 report the results of the simulations.

Bibliography

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

Davis, R. A., W. Dunsmuir, and Y. Wang (1999). Modelling time series of count data. In S. Ghosh (Ed.), *Asymptotics, Nonparametrics and Time Series*, pp. 63–113. New York: Marcel Dekker.

Diaconis, P. and D. Freedman (1999). Iterated random functions. *SIAM Review 41*(1), 45–76.

Duflo, M. (1997). *Random Iterative Models*, Volume 34 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. Springer.

Francq, C. and J.-M. Zakoian (2004). Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli 10*(4), 605–637.

Ghalanos, A. and S. Theussl (2012). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.14.

Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance 14*(4), 564–583.

Table C.3: Simulation from Model I with $\boldsymbol{\theta} = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 0.3)^T$.

n		δ_1	δ_2	α_1	α_2	β_{11}	β_{12}	β_{21}	β_{22}	φ
50	bias	0.6413	0.1340	−0.0203	0.0444	−0.0549	0.0108	−0.0088	−0.0684	0.1374
	std	1.5963	0.6533	0.1909	0.1935	0.1342	0.1881	0.0627	0.1340	0.4674
	rmse	1.7195	0.6666	0.1919	0.1984	0.1449	0.1883	0.0633	0.1504	0.4870
100	bias	0.4163	0.0317	−0.0232	0.0239	−0.0276	0.0070	−0.0019	−0.0292	0.0575
	std	1.2095	0.4767	0.1553	0.1571	0.0941	0.1524	0.0460	0.1007	0.3574
	rmse	1.2786	0.4775	0.1570	0.1588	0.0980	0.1525	0.0460	0.1048	0.3618
250	bias	0.1929	−0.0021	−0.0173	0.0134	−0.0060	0.0023	−0.0011	−0.0103	0.0240
	std	0.8410	0.3320	0.1219	0.1226	0.0626	0.1049	0.0317	0.0662	0.2519
	rmse	0.8624	0.3319	0.1230	0.1233	0.0628	0.1049	0.0317	0.0670	0.2529
500	bias	0.1285	0.0018	−0.0117	0.0075	−0.0034	0.0013	−0.0004	−0.0061	0.0139
	std	0.6087	0.2412	0.0887	0.0989	0.0471	0.0743	0.0230	0.0450	0.1986
	rmse	0.6219	0.2411	0.0894	0.0991	0.0471	0.0743	0.0230	0.0454	0.1990
1000	bias	0.0275	0.0093	−0.0008	−0.0002	−0.0026	−0.0008	−0.0003	−0.0028	−0.0002
	std	0.4168	0.1812	0.0633	0.0734	0.0305	0.0506	0.0163	0.0334	0.1449
	rmse	0.4175	0.1814	0.0633	0.0734	0.0306	0.0506	0.0163	0.0335	0.1448

Table C.4: Simulation from Model I with $\boldsymbol{\theta} = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 0.3)^T$. For each model, an entry in the table reports the fraction of the 1000 runs for which the given measure has selected the specific model as the best or the worst fit, respectively.

		Best fit					Worst fit				
n	Model	AIC	BIC	LS	QS	RPS	AIC	BIC	LS	QS	RPS
50	I	0.068	0.014	0.370	0.303	0.298	0.337	0.481	0.019	0.053	0.016
	II	0.587	0.891	0.050	0.100	0.044	0.308	0.052	0.916	0.812	0.900
	III	0.204	0.052	0.271	0.299	0.318	0.214	0.286	0.043	0.082	0.057
	IV	0.141	0.043	0.309	0.298	0.340	0.141	0.181	0.022	0.053	0.027
100	I	0.183	0.035	0.413	0.307	0.330	0.191	0.401	0.001	0.031	0.003
	II	0.286	0.794	0.005	0.062	0.010	0.610	0.134	0.988	0.911	0.980
	III	0.272	0.077	0.265	0.338	0.339	0.129	0.303	0.005	0.031	0.007
	IV	0.259	0.094	0.317	0.293	0.321	0.070	0.162	0.006	0.027	0.010
250	I	0.343	0.215	0.390	0.326	0.317	0.011	0.182	0	0.003	0
	II	0.022	0.413	0	0.011	0	0.965	0.505	0.999	0.987	0.999
	III	0.336	0.198	0.311	0.368	0.378	0.015	0.220	0	0.005	0
	IV	0.299	0.174	0.299	0.295	0.305	0.009	0.093	0.001	0.005	0.001
500	I	0.438	0.413	0.331	0.357	0.339	0	0.019	0	0	0
	II	0	0.064	0	0.003	0	0.999	0.905	1	0.997	1
	III	0.284	0.269	0.382	0.345	0.355	0	0.056	0	0.003	0
	IV	0.278	0.254	0.287	0.295	0.306	0.001	0.020	0	0	0
1000	I	0.561	0.561	0.261	0.359	0.370	0	0	0	0	0
	II	0	0	0	0	0	1	0.999	1	1	1
	III	0.186	0.186	0.401	0.372	0.363	0	0.001	0	0	0
	IV	0.253	0.253	0.338	0.269	0.267	0	0	0	0	0

Table C.5: Simulation from Model II with $\boldsymbol{\theta} = (3, 0.1, 0.2, 2, 0.4, 0.2)^T$.

n		δ_1	α_1	β_1	δ_2	α_2	β_2	φ_{I}	φ_{III}	φ_{IV}
50	bias	−0.0581	0.0437	−0.0268	1.0242	−0.1826	−0.0210	0.2980	0.0085	0.0570
	std	1.1027	0.2248	0.1201	1.4741	0.2602	0.1267	0.4000	0.1941	0.9391
	rmse	1.1037	0.2289	0.1230	1.7944	0.3178	0.1283	0.4987	0.1942	0.9403
100	bias	−0.1276	0.0518	−0.0199	0.5620	−0.1058	−0.0096	0.2107	0.0078	0.0305
	std	0.9711	0.2216	0.0934	1.2741	0.2591	0.0942	0.2963	0.1294	0.6636
	rmse	0.9790	0.2275	0.0955	1.3920	0.2798	0.0946	0.3634	0.1295	0.6640
250	bias	−0.1789	0.0497	−0.0077	0.2880	−0.0544	−0.0036	0.1142	0.0016	−0.0125
	std	0.8066	0.1953	0.0642	0.9982	0.2185	0.0607	0.1773	0.0819	0.4008
	rmse	0.8258	0.2015	0.0646	1.0384	0.2251	0.0608	0.2109	0.0818	0.4008
500	bias	−0.1451	0.0386	−0.0053	0.1102	−0.0193	−0.0029	0.0890	0.0007	0.0010
	std	0.6343	0.1555	0.0435	0.7305	0.1630	0.0440	0.1254	0.0501	0.2838
	rmse	0.6503	0.1601	0.0438	0.7384	0.1640	0.0440	0.1537	0.0501	0.2836
1000	bias	−0.0635	0.0169	−0.0022	0.0763	−0.0157	0.0003	0.0592	−0.0017	−0.0081
	std	0.4856	0.1178	0.0318	0.5011	0.1139	0.0306	0.0878	0.0347	0.2067
	rmse	0.4895	0.1189	0.0319	0.5066	0.1150	0.0305	0.1058	0.0347	0.2068

Table C.6: Simulation from Model II with $\boldsymbol{\theta} = (3, 0.1, 0.2, 2, 0.4, 0.2)^T$. For each model, an entry in the table reports the fraction of the 1000 runs for which the given measure has selected the specific model as the best or the worst fit, respectively.

		Best fit					Worst fit				
n	Model	AIC	BIC	LS	QS	RPS	AIC	BIC	LS	QS	RPS
50	I	0.018	0.001	0.394	0.260	0.302	0.553	0.572	0.071	0.110	0.075
	II	0.898	0.994	0.257	0.249	0.208	0.034	0.002	0.618	0.524	0.606
	III	0.054	0.004	0.174	0.252	0.241	0.259	0.267	0.189	0.205	0.165
	IV	0.030	0.001	0.175	0.239	0.249	0.154	0.159	0.122	0.161	0.154
100	I	0.009	0	0.402	0.191	0.238	0.582	0.600	0.060	0.128	0.104
	II	0.900	0.999	0.286	0.252	0.248	0.029	0	0.650	0.546	0.613
	III	0.050	0.001	0.174	0.288	0.275	0.236	0.240	0.168	0.193	0.171
	IV	0.041	0	0.138	0.269	0.239	0.153	0.160	0.122	0.133	0.112
250	I	0.017	0.001	0.440	0.185	0.215	0.627	0.647	0.063	0.138	0.122
	II	0.908	0.998	0.349	0.344	0.337	0.034	0.001	0.609	0.488	0.564
	III	0.046	0.001	0.104	0.254	0.233	0.173	0.181	0.180	0.189	0.162
	IV	0.029	0	0.107	0.217	0.215	0.166	0.171	0.148	0.185	0.152
500	I	0.009	0	0.455	0.196	0.231	0.594	0.610	0.044	0.152	0.112
	II	0.920	0.999	0.326	0.376	0.331	0.024	0	0.663	0.511	0.616
	III	0.038	0	0.136	0.228	0.222	0.213	0.217	0.153	0.161	0.144
	IV	0.033	0.001	0.083	0.200	0.216	0.169	0.173	0.140	0.176	0.128
1000	I	0.018	0	0.470	0.180	0.223	0.615	0.627	0.046	0.161	0.115
	II	0.894	1	0.353	0.423	0.363	0.034	0	0.631	0.489	0.591
	III	0.045	0	0.114	0.206	0.219	0.198	0.210	0.163	0.192	0.147
	IV	0.043	0	0.063	0.191	0.195	0.153	0.163	0.160	0.158	0.147

Table C.7: Simulation from Model III with $\theta = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 3)^T$.

n		δ_1	δ_2	α_1	α_2	β_{11}	β_{12}	β_{21}	β_{22}	φ
50	bias	0.2863	0.0466	0.0103	0.0265	−0.0470	0.0036	−0.0006	−0.0520	0.5112
	std	1.2838	0.5551	0.1682	0.1742	0.1320	0.1798	0.0721	0.1354	0.9202
	rmse	1.3147	0.5568	0.1684	0.1762	0.1400	0.1798	0.0720	0.1450	1.0522
100	bias	0.1499	0.0366	0.0031	0.0047	−0.0285	0.0144	−0.0019	−0.0195	0.2323
	std	0.9249	0.4064	0.1248	0.1271	0.1015	0.1564	0.0598	0.1091	0.5406
	rmse	0.9365	0.4078	0.1248	0.1271	0.1053	0.1570	0.0598	0.1108	0.5881
250	bias	0.1285	0.0409	−0.0059	−0.0042	−0.0105	0.0022	−0.0027	−0.0052	0.0973
	std	0.6336	0.2715	0.0871	0.0910	0.0745	0.1180	0.0417	0.0755	0.3203
	rmse	0.6462	0.2744	0.0872	0.0911	0.0752	0.1180	0.0417	0.0757	0.3346
500	bias	0.0558	0.0217	−0.0041	−0.0065	−0.0029	−0.0008	−0.0001	−0.0029	0.0505
	std	0.4295	0.1924	0.0615	0.0685	0.0530	0.0920	0.0312	0.0583	0.2306
	rmse	0.4329	0.1936	0.0616	0.0688	0.0531	0.0920	0.0312	0.0583	0.2360
1000	bias	0.0292	0.0156	0.0003	−0.0032	−0.0042	0.0007	−0.0007	−0.0005	0.0212
	std	0.3070	0.1405	0.0430	0.0526	0.0370	0.0646	0.0209	0.0398	0.1610
	rmse	0.3083	0.1413	0.0430	0.0526	0.0372	0.0646	0.0209	0.0398	0.1623

Table C.8: Simulation from Model III with $\theta = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 3)^T$. For each model, an entry in the table reports the fraction of the 1000 runs for which the given measure has selected the specific model as the best or the worst fit, respectively.

		Best fit					Worst fit				
n	Model	AIC	BIC	LS	QS	RPS	AIC	BIC	LS	QS	RPS
50	I	0.001	0.001	0.005	0.041	0.027	0.001	0.016	0.825	0.753	0.697
	II	0	0	0.341	0.337	0.329	0.997	0.982	0.065	0.047	0.121
	III	0.923	0.923	0.495	0.365	0.449	0.001	0.001	0.029	0.048	0.053
	IV	0.076	0.076	0.159	0.257	0.195	0.001	0.001	0.081	0.152	0.129
100	I	0	0	0	0.015	0.004	0	0	0.948	0.879	0.845
	II	0	0	0.182	0.245	0.195	1	1	0.029	0.016	0.093
	III	0.988	0.988	0.741	0.522	0.672	0	0	0	0.017	0.010
	IV	0.012	0.012	0.077	0.218	0.129	0	0	0.023	0.088	0.052
250	I	0	0	0	0	0	0	0	0.997	0.991	0.972
	II	0	0	0.042	0.113	0.063	1	1	0.003	0	0.027
	III	1	1	0.956	0.770	0.916	0	0	0	0	0
	IV	0	0	0.002	0.117	0.021	0	0	0	0.009	0.001
500	I	0	0	0	0	0	0	0	1	0.998	0.997
	II	0	0	0.002	0.028	0.004	1	1	0	0	0.003
	III	1	1	0.998	0.940	0.994	0	0	0	0	0
	IV	0	0	0	0.032	0.002	0	0	0	0.002	0
1000	I	0	0	0	0	0	0	0	1	1	1
	II	0	0	0	0.001	0	1	1	0	0	0
	III	1	1	1	0.992	1	0	0	0	0	0
	IV	0	0	0	0.007	0	0	0	0	0	0

Table C.9: Simulation from Model IV with $\boldsymbol{\theta} = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 3)^T$.

n		δ_1	δ_2	α_1	α_2	β_{11}	β_{12}	β_{21}	β_{22}	φ
50	bias	0.5304	0.0581	−0.0163	0.0279	−0.0570	0.0281	0.0021	−0.0622	0.2918
	std	1.5095	0.6325	0.1870	0.1849	0.1375	0.1952	0.0679	0.1333	1.0814
	rmse	1.5992	0.6348	0.1876	0.1869	0.1488	0.1971	0.0679	0.1471	1.1196
100	bias	0.3183	0.0085	−0.0079	0.0250	−0.0323	0.0110	−0.0009	−0.0284	0.1541
	std	1.1854	0.4433	0.1577	0.1567	0.1029	0.1545	0.0539	0.1038	0.7083
	rmse	1.2780	0.4549	0.1630	0.1621	0.1128	0.1581	0.0560	0.1116	2.3505
250	bias	0.1345	0.0242	−0.0076	0.0042	−0.0085	0.0005	−0.0016	−0.0094	0.0400
	std	0.7655	0.2995	0.1102	0.1108	0.0670	0.1085	0.0353	0.0690	0.4302
	rmse	0.7769	0.3004	0.1104	0.1109	0.0675	0.1084	0.0353	0.0696	0.4319
500	bias	0.0645	0.0144	−0.0040	−0.0012	−0.0057	0.0072	−0.0007	−0.0026	0.0261
	std	0.5466	0.2276	0.0795	0.0844	0.0495	0.0817	0.0247	0.0475	0.3039
	rmse	0.5501	0.2280	0.0796	0.0844	0.0498	0.0820	0.0247	0.0476	0.3049
1000	bias	0.0503	0.0153	−0.0037	−0.0032	−0.0012	−0.0026	0.0000	−0.0023	0.0119
	std	0.3781	0.1603	0.0558	0.0654	0.0335	0.0560	0.0178	0.0337	0.2182
	rmse	0.3812	0.1610	0.0559	0.0655	0.0335	0.0560	0.0178	0.0338	0.2184

Table C.10: Simulation from Model IV with $\boldsymbol{\theta} = (3, 0.8, 0.2, 0.1, 0.4, 0.2, 0.1, 0.3, 3)^T$. For each model, an entry in the table reports the fraction of the 1000 runs for which the given measure has selected the specific model as the best or the worst fit, respectively.

		Best fit					Worst fit				
n	Model	AIC	BIC	LS	QS	RPS	AIC	BIC	LS	QS	RPS
50	I	0.055	0.026	0.104	0.213	0.177	0.047	0.292	0.152	0.193	0.127
	II	0.041	0.285	0.079	0.108	0.075	0.844	0.368	0.766	0.632	0.752
	III	0.209	0.153	0.408	0.273	0.316	0.094	0.304	0.051	0.117	0.083
	IV	0.695	0.536	0.409	0.406	0.432	0.015	0.036	0.031	0.058	0.038
100	I	0.041	0.031	0.050	0.163	0.083	0.002	0.048	0.153	0.207	0.120
	II	0	0.037	0.024	0.068	0.029	0.986	0.809	0.830	0.681	0.842
	III	0.117	0.114	0.429	0.301	0.382	0.012	0.140	0.009	0.068	0.025
	IV	0.842	0.818	0.497	0.468	0.506	0	0.003	0.008	0.044	0.013
250	I	0.005	0.005	0.009	0.088	0.024	0	0	0.063	0.132	0.049
	II	0	0	0	0.012	0	1	1	0.937	0.844	0.949
	III	0.026	0.026	0.399	0.356	0.370	0	0	0	0.015	0.002
	IV	0.969	0.969	0.592	0.544	0.606	0	0	0	0.009	0
500	I	0	0	0	0.032	0.002	0	0	0.015	0.064	0.016
	II	0	0	0	0.004	0	1	1	0.985	0.933	0.984
	III	0.001	0.001	0.272	0.334	0.327	0	0	0	0.003	0
	IV	0.999	0.999	0.728	0.630	0.671	0	0	0	0	0
1000	I	0	0	0	0.005	0	0	0	0.006	0.028	0.005
	II	0	0	0	0	0	1	1	0.994	0.972	0.995
	III	0.001	0.001	0.152	0.303	0.255	0	0	0	0	0
	IV	0.999	0.999	0.848	0.692	0.745	0	0	0	0	0

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2012). *copula: Multivariate Dependence with Copulas*. R package version 0.999-5.

Hofert, M. and M. Maechler (2011). Nested archimedean copulas meet R: The nacopula package. *Journal of Statistical Software 39*(9), 1–20.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. Wiley New York.

Jung, R. C. and A. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling 6*(2), 81–96.

Jung, R. C. and A. R. Tremayne (2011). Useful models for time series of counts or simply wrong ones? *Advances in Statistical Analysis 95*(1), 59–91.

Kocherlakota, S. and K. Kocherlakota (1992). *Bivariate Discrete Distributions*, Volume 132 of *STATISTICS: textbooks and monographs*. New York: Marcel Dekker.

Kojadinovic, I. and J. Yan (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software 34*(9), 1–20.

Liu, H. (2012). *Some Models for Time Series of Counts*. Ph. D. thesis, Columbia University.

Mardia, K. V. (1970). *Families of Bivariate Distributions*, Volume 27 of *Griffin's Statistical Monographs and Courses*. Griffin London.

McKenzie, E. (2003). Discrete variate time series. *Handbook of statistics 21*, 573–606.

Meyn, S. P. and R. L. Tweedie (2009). *Markov Chains and Stochastic Stability* (2nd ed.). Cambridge University Press.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer.

Pedeli, X. and D. Karlis (2011). A bivariate INAR (1) process with application. *Statistical modelling 11*(4), 325–349.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris. 8*, 229–231.

Wang, C., H. Liu, J.-F. Yao, R. Davis, and W. Li (2012). Self-excited threshold Poisson autoregression. Submitted.

Wu, W. B. and X. Shao (2004). Limit theorems for iterated random functions. *Journal of Applied Probability 41*(2), 425–436.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software 21*(4), 1–21.

Ye, Y. (1987). *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. Ph. D. thesis, Department of EES, Stanford University.