Journal of Multivariate Analysis 173 (2019) 181–203

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Bivariate integer-autoregressive process with an application to mutual fund flows

Serge Darolles [a], Gaëlle Le Fol [a], Yang Lu [b,*], Ran Sun [a]

[a] Université Paris-Dauphine, PSL Research University, CNRS, DRM, Finance, 75016 Paris, France
[b] Université Paris 13, Sorbonne Paris Cité, CNRS, CEPN, 93430, Villetaneuse, France

ARTICLE INFO

Article history:
Received 15 August 2018
Available online 28 February 2019

JEL classification:
C32
C53

AMS 2010 subject classifications:
primary 60G10
secondary 60E05

Keywords:
Compound autoregressive process
Memory persistence
Mutual funds
Non-linear forecasting

ABSTRACT

We propose a new family of bivariate nonnegative integer-autoregressive (BINAR) models for count process data. We first generalize the existing BINAR(1) model by allowing for dependent thinning operators and arbitrary innovation distribution. The extended family allows for intuitive interpretation, as well as tractable aggregation and stationarity properties. We then introduce higher order BINAR(p) and BINAR(∞) dynamics to accommodate more flexible serial dependence patterns. So far, the literature has regarded such models as computationally intractable. We show that the extended BINAR family allows for closed-form predictive distributions at any horizons and for any values of p, which significantly facilitates non-linear forecasting and likelihood based estimation. Finally, a BINAR(∞) model with memory persistence is applied to open-ended mutual fund purchase and redemption order counts.

1. Introduction

Nonnegative low count processes have been widely used in domains such as marketing [4], economics [3], finance [22], insurance [20] and beyond, ever since the seminal work of McKenzie [34]. Our interest in this paper lies in the monitoring of the liquidity risk of an open-ended mutual fund (MF). An MF channels investors' cash investment into less liquid assets, and is thus structurally vulnerable to liquidity risk. This risk has recently received much attention from the regulators [1,10], but its quantification and management remain difficult. Indeed, from the modeling point of view, the liquidity risk is quite different from traditional market risks in that it involves the daily counts of redemption and purchase orders, which are (i) most of the time low integers or zero, but also have a non-null probability of taking mildly large values; (ii) both cross-sectionally and serially dependent, with significant heteroscedasticity.

Recently, the MF industry has started to record purchase and redemption order count data separately. This allows to distinguish auto-correlation effects and cross-effects between the two count processes, which have different economic interpretations. For instance, the clustering of the redemption counts corresponds to fund run, whereas a fund manager usually reacts to past redemptions by seeking new investors in order to stabilize the fund size, leading to a positive feedback effect between past redemption and current purchase counts. Therefore, a bivariate count analysis can be of great interest to understand clients' behavior and the manager's reaction to exogenous liquidity shock.

* Corresponding author.
E-mail address: yang.lu@univ-paris13.fr (Y. Lu).

https://doi.org/10.1016/j.jmva.2019.02.015

Yet the literature on bivariate count processes is still in its infancy. The benchmark approach is the first order Bivariate INteger-valued AutoRegressive model [27,36], which assumes that, for each time t,

$$\begin{aligned} X_{1,t} &= \alpha_{11} \circ X_{1,t-1} + \alpha_{12} \circ X_{2,t-1} + \epsilon_{1,t}, \\ X_{2,t} &= \alpha_{21} \circ X_{1,t-1} + \alpha_{22} \circ X_{2,t-1} + \epsilon_{2,t}, \end{aligned} \tag{1}$$

where given $\boldsymbol{X}_{t-1} = (X_{1,t-1}, X_{2,t-1})^\top$, the binomial thinning operators are defined as follows: for each $i, j \in \{1, 2\}$, the variable $\alpha_{i,j} \circ X_{j,t-1}$ is binomial with size $X_{j,t-1}$ and success probability $\alpha_{i,j} \in [0, 1]$. Moreover, these variables are conditionally independent, and are also independent of the i.i.d. innovation sequence $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \epsilon_{2,t})^\top$.

This approach has several drawbacks. First, the conditional independence assumption between the thinning operators restricts significantly the dependence feature. Second, so far only Latour [27] has considered higher-order models, but he suggests to base the estimation and forecasting solely on conditional expectation, i.e., as if the observations are continuous, and nothing is said about the empirical performance of this approach. This is due to the fact that the term structure of predictive distributions of higher-order BINAR process is yet to be derived and is so far (wrongly) considered intractable. These downsides seriously limit their usefulness for risk management and forecasting purposes.

Besides BINAR processes, other non-thinning-based models have been introduced. Quoreshi [37] and Livsey et al. [29] proposed parameter-driven models with flexible (auto-)correlation, but the likelihood estimation and forecasting in these models are way too cumbersome to be feasible. Another popular approach is the bivariate INGARCH model [28], which assumes that given the past, $X_{1,t}$ and $X_{2,t}$ follow some simple (say, Poisson) distributions, with parameters that are exponentially weighted averages of past observations. Then the contemporaneous conditional dependence between $X_{1,t}$ and $X_{2,t}$ is captured by a copula [2,13,22]. The downsides of the latter approach are that (i) it is sometimes unclear whether such processes are strictly stationary and how memory persistence can be introduced; (ii) forecasting formulas beyond the conditional expectation necessitate numerical integration; (iii) it remains an open question as to whether the copula is identified in the count data setting, as discussed, e.g., in [16,17]. Recently, some alternative, non-INGARCH models were proposed in [9,21,40], and they allow for rather flexible serial dependence. However, it is computationally difficult to extend them to models of arbitrary orders. For instance, the model of [9], as well as INGARCH models, are necessarily infinite-order Markov, whereas that of [21] is first-order Markov. Further, [9,21] assume ex ante the conditional distributions of $X_{1,t}$ and $X_{2,t}$ given the past to be equi- (resp. over-) dispersed.

The paper that is closest to ours is that of Scotto et al. [40]. While these authors are mainly interested in bounded counts, they mention in their conclusion some possible extensions of model (1), which they conjecture to be appropriate for unbounded count data. In this paper, we show that one of these extensions has indeed tractable properties, even after extension to higher-order cases.

More precisely, we contribute to the BINAR literature in two ways. First, we extend model (1), called independent BINAR(1) henceforth, by introducing (positively or negatively) dependent thinning operators and arbitrary innovation distribution. We show that the process belongs to the compound autoregressive (CaR) family, and possesses intuitive aggregation and stationarity properties. We then go on to clarify that in this family of BINAR models, the predictive distributions at various horizons are easily computable via a matrix-based algorithm. This largely facilitates likelihood-based inference and non-linear forecasting, especially when it comes to the prediction of extreme events. Second, we extend our model to higher-order dependent BINAR(p) and BINAR(∞) processes, which can better capture slowly decaying serial correlation patterns.

The paper is organized as follows. The dependent BINAR(1) model is introduced in Section 2 and extended in Section 3 to higher-order BINAR(p) and BINAR(∞) models. The predictive distributions are computed in Section 4 and the model is applied in Section 5 to forecast the counts of share purchase and redemption of an MF. Section 6 contains concluding remarks. Proofs and technical details are gathered in Appendix A.

2. Dependent BINAR(1) process

2.1. The dynamic specification

Definition 1. We say that the bivariate count process $\boldsymbol{X}_t = (X_{1,t}, X_{2,t})^\top$ with domain $\mathbb{N}^2 = \{0, 1, \ldots\}^2$ is dependent BINAR(1) if it has the stochastic representation

$$\forall_t \quad \begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \sum_{i=1}^{X_{1,t-1}} \begin{bmatrix} Z_{1,i,t} \\ Z_{2,i,t} \end{bmatrix} + \sum_{j=1}^{X_{2,t-1}} \begin{bmatrix} Z_{3,j,t} \\ Z_{4,j,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}, \tag{2}$$

where, given $\boldsymbol{X}_{t-1}$,

(i) for i, j, t varying, random vectors $(Z_{1,i,t}, Z_{2,i,t})^\top$, $(Z_{3,j,t}, Z_{4,j,t})^\top$, $(\epsilon_{1,t}, \epsilon_{2,t})^\top$ are mutually independent copies of $(Z_1, Z_2)^\top$, $(Z_3, Z_4)^\top$, and $(\epsilon_1, \epsilon_2)^\top$, respectively.

(ii) Couples $(Z_1, Z_2)^\top$ and $(Z_3, Z_4)^\top$ are mutually independent and bivariate Bernoulli distributed. That is, they have Bernoulli marginal distribution with success probabilities, viz.

$$\begin{bmatrix} \Pr(Z_1 = 1) & \Pr(Z_3 = 1) \\ \Pr(Z_2 = 1) & \Pr(Z_4 = 1) \end{bmatrix} = \begin{bmatrix} \alpha_{11}, & \alpha_{12} \\ \alpha_{21}, & \alpha_{22} \end{bmatrix} = A,$$

where all the entries of the matrix A are nonnegative, whereas the two joint probabilities are respectively

$$\Pr(Z_1 = Z_2 = 1) = q_1, \quad \Pr(Z_3 = Z_4 = 1) = q_2.$$

(iii) The innovations $\boldsymbol{\epsilon}_t$ are independent of $\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots$, and i.i.d. across t. Moreover, they are nonnegative and have finite variance.

The bivariate Bernoulli distribution is first introduced in [42] (see also [31]), and is recently used in [40] to model bounded count processes. In order for it to be well defined, its parameters have to satisfy the following constraint.

Lemma 1 *([23], p. 210)*. *The parameters $(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, q_1, q_2)$ of the bivariate Bernoulli distribution can take any values satisfying the following two inequalities:*

$$\max(\alpha_{11} + \alpha_{21} - 1, 0) \le q_1 \le \min(\alpha_{11}, \alpha_{21}), \tag{3}$$

$$\max(\alpha_{12} + \alpha_{22} - 1, 0) \le q_2 \le \min(\alpha_{12}, \alpha_{22}). \tag{4}$$

Note that the covariance between Z_1 and Z_2 (resp. Z_3 and Z_4) is $q_1 - \alpha_{11}\alpha_{21}$. In particular, if $q_1 = \max(\alpha_{11} + \alpha_{21} - 1, 0)$, then the covariance is nonpositive; if $q_1 = \min(\alpha_{11}, \alpha_{21})$, the covariance is nonnegative; if $q_1 = \alpha_{11}\alpha_{21}$ and $q_2 = \alpha_{12}\alpha_{22}$, we get the independent BINAR(1) model (1).

Birth–death-immigration interpretation. Let us temporarily assume that $X_{1,t}$ and $X_{2,t}$ count individuals of type 1 and type 2 at time t, respectively. A type $j_1 \in \{1, 2\}$ individual produces a type $j_2 \in \{1, 2\}$ off-spring with marginal probability α_{j_1,j_2}, and joint probability q_{j_1}. Then the population of type $j \in \{1, 2\}$ at time $t + 1$ is composed of such off-springs, plus $\epsilon_{j,t+1}$ immigrants of type j. In particular, if $q_1 = \alpha_{11}\alpha_{21}$ and $q_2 = \alpha_{12}\alpha_{22}$, the productions of the two types of off-springs are independent events. If instead $q_1 = q_2 = 0$, the productions are mutually exclusive, i.e., each individual can only produce up to one off-spring.

2.2. Conditional distribution

2.2.1. First two conditional moments

Since we have

$$\mathrm{E}(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}) = A\boldsymbol{X}_{t-1} + \mathrm{E}(\boldsymbol{\epsilon}_t), \tag{5}$$

we abbreviate model (2) into

$$\boldsymbol{X}_t = A(q_1, q_2) \circ \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t,$$

where the operator $A(q_1, q_2)\circ$ is linear and will be called the dependent matrix thinning operator. The conditional covariance matrix is

$$\begin{bmatrix} \mathrm{var}(X_{1,t}|\boldsymbol{X}_{t-1}) & * \\ \mathrm{cov}(X_{1,t}, X_{2,t}|\boldsymbol{X}_{t-1}) & \mathrm{var}(X_{2,t}|\boldsymbol{X}_{t-1}) \end{bmatrix} = \gamma_\epsilon + X_{1,t-1}\gamma_{12} + X_{2,t-1}\gamma_{34},$$

where γ_ϵ, γ_{12}, γ_{34} are the covariance matrices of $(\epsilon_{1,t}, \epsilon_{2,t})$, $(Z_{1,t}, Z_{2,t})$ and $(Z_{3,t}, Z_{4,t})$, respectively, viz.

$$\gamma_{12} = \begin{bmatrix} \alpha_{11}(1 - \alpha_{11}) & * \\ q_1 - \alpha_{11}\alpha_{21} & \alpha_{21}(1 - \alpha_{21}) \end{bmatrix}, \quad \gamma_{34} = \begin{bmatrix} \alpha_{12}(1 - \alpha_{12}) & * \\ q_2 - \alpha_{12}\alpha_{22} & \alpha_{22}(1 - \alpha_{22}) \end{bmatrix}, \tag{6}$$

and the conditional correlation between $X_{1,t}$ and $X_{2,t}$ is

$$\mathrm{corr}(X_{1,t}, X_{2,t}|\boldsymbol{X}_{t-1}) = \frac{\mathrm{cov}(\epsilon_{1,t}, \epsilon_{2,t}) + X_{1,t-1}\mathrm{cov}(Z_1, Z_2) + X_{2,t-1}\mathrm{cov}(Z_3, Z_4)}{\sqrt{\left\{\mathrm{var}(\epsilon_{1,t}) + X_{1,t-1}\mathrm{var}(Z_1) + X_{2,t-1}\mathrm{var}(Z_3)\right\}\left\{\mathrm{var}(\epsilon_{2,t}) + X_{1,t-1}\mathrm{var}(Z_2) + X_{2,t-1}\mathrm{var}(Z_4)\right\}}}.$$

In the above equation, the denominator of the right-hand side does not depend on q_1 or q_2, whereas the numerator is increasing in q_1 and q_2. In the independent BINAR(1) model with $q_1 = \alpha_{11}\alpha_{21}$, $q_2 = \alpha_{22}\alpha_{12}$, the above correlation coefficient is (in absolute value) no larger than $\mathrm{corr}(\epsilon_{1,t}, \epsilon_{2,t})$, and becomes small whenever $X_{1,t-1}$ and/or $X_{2,t-1}$ are large. Thus, the conditional heteroscedasticity cannot be well captured. This downside exists also in several other competing bivariate count process models. For instance, in the bivariate Poisson autoregression of [28], the conditional correlation coefficient goes to zero when $X_{1,t}$ and $X_{2,t}$ are large. In copula-based bivariate count processes [22], this conditional correlation coefficient can only be computed numerically and it is not clear how it behaves when components of $\boldsymbol{X}_{t-1}$ are large; in our model, however, when both $X_{1,t-1}$ and $X_{2,t-1}$ are large, the conditional correlation is approximately

$$\mathrm{corr}(X_{1,t}, X_{2,t}|\boldsymbol{X}_{t-1}) \approx \frac{X_{1,t-1}\mathrm{cov}(Z_1, Z_2) + X_{2,t-1}\mathrm{cov}(Z_3, Z_4)}{\sqrt{\left\{X_{1,t-1}\mathrm{var}(Z_1) + X_{2,t-1}\mathrm{var}(Z_3)\right\}\left\{X_{1,t-1}\mathrm{var}(Z_2) + X_{2,t-1}\mathrm{var}(Z_4)\right\}}},$$

which can be closer to 1 (resp. -1) if q_1 and q_2 are close to their upper (resp. lower) bounds.

Finally, if $\epsilon_{1,t}$, $\epsilon_{2,t}$ have Poisson distributions, then both components are conditionally under-dispersed, viz.

$$\forall_{j\in\{1,2\}} \quad \mathrm{var}(X_{jt}|\boldsymbol{X}_{t-1}) \leq \mathrm{E}(X_{jt}|\boldsymbol{X}_{t-1}).$$

This differs from the models of [9,21], which are conditionally equi-dispersed and over-dispersed, respectively. In general, by leaving the distribution of $\boldsymbol{\epsilon}_t$ unconstrained, the BINAR(1) family allows for flexible conditional dispersion.

2.2.2. Conditional pgf

The dynamics of the process $\boldsymbol{X}_t$ is characterized by its conditional probability generating function (pgf), which is given, for any $u, v \geq 0$, by

$$\begin{aligned}\mathrm{E}(u^{X_{1,t}}v^{X_{2,t}}|\boldsymbol{X}_{t-1}) &= \left\{\mathrm{E}(u^{Z_1}v^{Z_2})\right\}^{X_{1,t-1}}\left\{\mathrm{E}(u^{Z_3}v^{Z_4})\right\}^{X_{2,t-1}}\mathrm{E}(u^{\epsilon_{1,t}}v^{\epsilon_{2,t}})\\ &= \{a_1(u,v)\}^{X_{1,t-1}}\{a_2(u,v)\}^{X_{2,t-1}}b(u,v),\end{aligned} \tag{7}$$

where $b(u, v)$ is the pgf of $(\epsilon_{1,t}, \epsilon_{2,t})$ and

$$\begin{aligned}a_1(u,v) &= q_1uv + (\alpha_{11} - q_1)u + (\alpha_{21} - q_1)v + (1 + q_1 - \alpha_{11} - \alpha_{21}),\\ a_2(u,v) &= q_2uv + (\alpha_{12} - q_2)u + (\alpha_{22} - q_2)v + (1 + q_2 - \alpha_{12} - \alpha_{22}),\end{aligned}$$

are the pgf of (Z_1, Z_2) and (Z_3, Z_4), respectively. This conditional pgf is exponentially affine in $\boldsymbol{X}_{t-1}$. Such a process is called compound autoregressive (CaR) [11]. A remarkable property of such processes is that the multiple-step-ahead conditional pgf remains exponentially affine in $\boldsymbol{X}_{t-1}$.

Proposition 1. *In model (2), we have, for any horizon $h \geq 0$ and any $u, v \geq 0$,*

$$\mathrm{E}(u^{X_{1,t+h-1}}v^{X_{2,t+h-1}}|\boldsymbol{X}_{t-1}) = \{a_1^{(h)}(u,v)\}^{X_{1,t-1}}\{a_2^{(h)}(u,v)\}^{X_{2,t-1}}b^{(h)}(u,v), \tag{8}$$

where functions $a_1^{(h)}(u,v)$, $a_2^{(h)}(u,v)$ and $b^{(h)}(u,v)$ are defined by the recursion:

$$\forall_{u,v\geq 0}\ \forall_{h\in\mathbb{N}} \quad \begin{aligned}a_1^{(h+1)}(u,v) &= 1 + \alpha_{11}\{a_1^{(h)}(u,v) - 1\} + \alpha_{21}\{a_2^{(h)}(u,v) - 1\} + q_1\{a_1^{(h)}(u,v) - 1\}\{a_2^{(h)}(u,v) - 1\},\\ a_2^{(h+1)}(u,v) &= 1 + \alpha_{12}\{a_1^{(h)}(u,v) - 1\} + \alpha_{22}\{a_2^{(h)}(u,v) - 1\} + q_2\{a_1^{(h)}(u,v) - 1\}\{a_2^{(h)}(u,v) - 1\},\\ b^{(h+1)}(u,v) &= b^{(h)}(u,v)b\{a_1^{(h)}(u,v), a_2^{(h)}(u,v)\},\end{aligned}$$

with initial conditions $a_1^{(0)}(u,v) = u$, $a_2^{(0)}(u,v) = v$, and $b^{(0)}(u,v) = 1$.

The proof (by induction) is straightforward and thus omitted. This proposition implies that for each $h \geq 1$, $a_1^{(h)}(u,v)$ and $a_2^{(h)}(u,v)$ are polynomials of degree 2^{h-1} in both arguments, except when $q_1 = q_2 = 0$. More precisely, we have the following result.

Corollary 1. *If $q_1 = q_2 = 0$, then*

$$\begin{bmatrix}a_1^{(h)}(u,v) - 1\\ a_2^{(h)}(u,v) - 1\end{bmatrix} = A^\top \begin{bmatrix}a_1^{(h-1)}(u,v) - 1\\ a_2^{(h-1)}(u,v) - 1\end{bmatrix} = (A^\top)^h \begin{bmatrix}u - 1\\ v - 1\end{bmatrix}. \tag{9}$$

Thus in this case $a_1^{(h)}(u,v)$ and $a_2^{(h)}(u,v)$ are affine instead of polynomial. In other words, the conditional pgf of $\{A(0,0)\circ\}^{(h)}\boldsymbol{X}_{t-1}$ given $\boldsymbol{X}_{t-1}$ at any horizon h has the same functional form as that of $A^h(0,0) \circ \boldsymbol{X}_{t-1}$, or equivalently $\{A(0,0)\circ\}^{(h)} = A^h(0,0)\circ$ is still a dependent matrix thinning operator, where $\{A(0,0)\circ\}^{(h)}$ is the h-time iteration of operator $A(0,0)\circ$. As a consequence, in terms of temporal aggregation, when observed at a lower frequency of h, the process $(X_{th})_t$ is still BINAR(1).

Remark 1. Corollary 1 is easily explained by the birth-immigration interpretation. If $q_1 = q_2 = 0$, each individual produces at most one off-spring in the next period, and thus at most one off-spring at horizon $h \geq 2$. Hence the identity $\{A(0,0)\circ\}^{(h)} = A^h(0,0)\circ$. Moreover, this constrained model has also a similar queuing interpretation as the univariate INAR(1) model [39]. More precisely, we can think of $X_{1,t}, X_{2,t}$ as the number of individuals in queues 1 and 2 at date t, respectively. Both queues have infinite capacity. At date $t+1$, $\epsilon_{1,t+1}$ (resp. $\epsilon_{2,t+1}$) new customers join queue 1 (resp. queue 2), whereas $X_{1,t}$ (resp. $X_{2,t}$) customers that were in queue 1 (resp. queue 2) at date t can either stay in the same queue, or go to the other queue, or leave the queues after being served, with probabilities α_{11}, α_{21} and $1 - \alpha_{11} - \alpha_{21}$ (resp. α_{22}, α_{12} and $1 - \alpha_{22} - \alpha_{12}$).

Remark 2. Note that the literature (see Theorem 2.8 in [5] as well as Eq. (15) in [36]) usually claims that in the independent BINAR model where $q_1 = \alpha_{11}\alpha_{21}$, $q_2 = \alpha_{21}\alpha_{22}$, the composite operator $\{A(q_1,q_2)\circ\}^{(h)}$ is equal to the bivariate thinning operator $A_h(q_{1,h}, q_{2,h})\circ$, where $A_h = A^h$ and $q_{1,h}$ (resp. $q_{2,h}$) is the products of the two entries of the first (resp. second) column of A^h. Then these authors deduce that the h-step-ahead conditional pgf still has the CaR form of Eq. (8), but with functions $a_1^{(h)}$ and $a_2^{(h)}$ being affine rather than higher-order polynomial. From the above analysis, we can see that this assertion is incorrect.

Notice also that Eq. (8) corresponds to the decomposition

$$\boldsymbol{X}_{t+h-1} = \{A(q_1,q_2)\circ\}^{(h)}\boldsymbol{X}_{t-1} + \{A(q_1,q_2)\circ\}^{(h-1)}\boldsymbol{\epsilon}_t + \cdots + A(q_1,q_2)\circ\boldsymbol{\epsilon}_{t+h-2} + \boldsymbol{\epsilon}_{t+h-1},$$

where the additive terms on the right-hand side are conditionally independent given $\boldsymbol{X}_{t-1}$, with conditional pgf's $a_1^{(h)}(u,v)^{X_{1,t-1}}a_2^{(h)}(u,v)^{X_{2,t-1}}$ for $\{A(q_1,q_2)\circ\}^{(h)}\boldsymbol{X}_{t-1}$, $b\{a_1^{(h-1)}(u,v), a_2^{(h-1)}(u,v)\}$ for $\{A(q_1,q_2)\circ\}^{(h-1)}\boldsymbol{\epsilon}_t,\ldots,$ and $b(u,v)$ for $\boldsymbol{\epsilon}_{t+h-1}$, respectively.

Example 1. Let us consider the case where $\epsilon_{1,t}$, $\epsilon_{2,t}$ are mutually independent, Poisson $\mathcal{P}(\lambda_1), \mathcal{P}(\lambda_2)$ distributed, respectively. We have $b(u,v) = \exp\{\lambda_1(u-1)+\lambda_2(v-1)\}$, and the pgf of $A(q_1,q_2)\circ\boldsymbol{\epsilon}_{t+h-2}$ is

$$\begin{aligned} b\{a_1(u,v), a_2(u,v)\} &= \exp\Big[\lambda_1\{a_1(u,v)-1\} + \lambda_2\{a_2(u,v)-1\}\Big] \\ &= \exp\Big[\lambda_1\big\{(\alpha_{11}-q_1)(u-1) + (\alpha_{21}-q_1)(v-1) + q_1(uv-1)\big\} \\ &\qquad + \lambda_2\big\{(\alpha_{12}-q_2)(u-1) + (\alpha_{22}-q_2)(v-1) + q_2(uv-1)\big\}\Big] \\ &= \exp\Big\{m_1(u-1) + m_2(v-1) + m_3(uv-1)\Big\}, \end{aligned} \tag{10}$$

where $m_1 = \lambda_1(\alpha_{11}-q_1)+\lambda_2(\alpha_{12}-q_2)$, $m_2 = \lambda_2(\alpha_{21}-q_1)+\lambda_2(\alpha_{22}-q_2)$, $m_3 = \lambda_1 q_1 + \lambda_2 q_2$. This is the pgf of a bivariate (dependent) Poisson distribution $\mathcal{BP}(m_1,m_2,m_3)$ with trivariate reduction [31] and its correlation coefficient is

$$\rho_1 = \frac{m_3}{\sqrt{(m_1+m_3)(m_2+m_3)}} = \frac{\lambda_1 q_1 + \lambda_2 q_2}{\sqrt{(\lambda_1\alpha_{11}+\lambda_2\alpha_{21})(\lambda_1\alpha_{12}+\lambda_2\alpha_{22})}},$$

which is nonnegative, and increasing in q_1 and q_2. As a consequence, the conditional distribution $\boldsymbol{X}_{t+1}|\boldsymbol{X}_{t-1}$ is $\mathcal{BP}(m_1 + \lambda_1, m_2+\lambda_2, m_3)$.

Under the same assumption, the pgf of $\{A(q_1,q_2)\circ\}^{(2)}\boldsymbol{\epsilon}_{t+h-3}$ is

$$b\{a_1^{(2)}(u,v), a_2^{(2)}(u,v)\} = \exp\Big[\lambda_1\{a_1^{(2)}(u,v)-1\} + \lambda_2\{a_2^{(2)}(u,v)-1\}\Big], \tag{11}$$

which is exponentially quadratic in u and v. The family of distributions with exponentially quadratic pgf is called bivariate Hermite (BH). It nests the bivariate Poisson as special case and is closed under convolution [25]. In particular, the conditional distribution $\boldsymbol{X}_{t+1}|\boldsymbol{X}_{t-1}$ still belongs to the BH family.

2.3. Cross-sectional aggregation

In the previous subsection, we have analyzed the frequency aggregation property of the BINAR(1) process. Brännäs et al. [6] also considered the cross-sectional aggregation of univariate INAR(1) models. Let us now extend this analysis to the dependent BINAR(1) models. Consider the sum process

$$X_{1,t} + X_{2,t} = \sum_{j=1}^{X_{1,t-1}}(Z_{1,j,t}+Z_{2,j,t}) + \sum_{j=1}^{X_{2,t-1}}(Z_{3,j,t}+Z_{4,j,t}) + \epsilon_{1,t} + \epsilon_{2,t}. \tag{12}$$

We can check that variables $Z_{1,j,t}+Z_{2,j,t}$ and $Z_{3,j,t}+Z_{4,j,t}$ are Bernoulli with parameters $\alpha_{11}+\alpha_{21}$ and $\alpha_{12}+\alpha_{22}$, respectively, if and only if they only take values 0 and 1, i.e., when $q_1 = q_2 = 0$. Nevertheless, given $X_{1,t-1}+X_{2,t-1}$, the sum

$$\sum_{j=1}^{X_{1,t-1}}(Z_{1,j,t}+Z_{2,j,t}) + \sum_{j=1}^{X_{2,t-1}}(Z_{3,j,t}+Z_{4,j,t})$$

is generically not binomial, except when $\alpha_{11}+\alpha_{21} = \alpha_{12}+\alpha_{22}$, in which case (12) becomes

$$X_{1,t} + X_{2,t} = \sum_{j=1}^{X_{1,t-1}+X_{2,t-1}}(Z_{1,j,t}+Z_{2,j,t}) + \epsilon_{1,t} + \epsilon_{2,t}, \tag{13}$$

where $Z_{1,j,t}+Z_{2,j,t}, j$ varying are independent of $\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots$ and the innovation $\epsilon_{1,t}+\epsilon_{2,t}$ is independent of the variables $Z_{1,j,t}+Z_{2,j,t}$. To summarize, we have the following property.

Proposition 2. *If in the BINAR(1) model $q_1 = q_2 = 0$ and $\alpha_{11}+\alpha_{21} = \alpha_{12}+\alpha_{22}$, then the sum process $X_{1,t}+X_{2,t}$ is also univariate INAR(1) with autocorrelation coefficient $\alpha_{11}+\alpha_{21}$.*

Thus the sum process is Markov with respect to its own history. Moreover, (13) says that $\ell(X_{1,t}+X_{2,t}|X_{1,t-1}, X_{2,t-1})$ depends on $X_{1,t-1}, X_{2,t-1}$ only via the sum $X_{1,t-1}+X_{2,t-1}$. In other words, the sum process can be viewed as an exogenous, common Markov factor.

We can interpret condition $\alpha_{11}+\alpha_{21}=\alpha_{12}+\alpha_{22}$ in terms of the second order dynamics of the process. To this end, we exclude two degenerate cases where the matrix A is diagonal or anti-diagonal. Then condition $\alpha_{11}+\alpha_{21}=\alpha_{12}+\alpha_{22}$ implies that

$$(1,1)^\top A=(\alpha_{11}+\alpha_{21})(1,1)^\top,$$

i.e., $\alpha_{11}+\alpha_{21}$ is an eigenvalue of the matrix A associated with the eigenvector $(1,1)^\top$. Since the matrix A and the vector $(1,1)^\top$ are positive, by the Perron–Frobenius theorem, $\alpha_{11}+\alpha_{21}$ is, in modulus, the simple largest eigenvalue of A. Thus, among all the linear combinations of components of $\boldsymbol{X}_t$, the process $X_{1,t}+X_{2,t}$ has the largest autocorrelation coefficient $\alpha_{11}+\alpha_{21}$. This justifies its interpretation as a common factor.

2.4. Stationarity

2.4.1. The stationarity condition

The strict stationarity condition of the BINAR(1) process is given in the next proposition.

Proposition 3. *Process $(\boldsymbol{X}_t)$ is strictly stationary if and only if*

$$(1-\alpha_{11})(1-\alpha_{22})>\alpha_{12}\alpha_{21}, \tag{14}$$

or equivalently, if and only if the eigenvalues of A are smaller than 1 in modulus.

Note that under condition (14), inequalities $\alpha_{21}+\alpha_{11}>1$ and $\alpha_{22}+\alpha_{12}>1$ cannot hold simultaneously. Thus in inequalities (3) and (4), at least one of the lower bounds is effectively zero.

2.4.2. The stationary distribution

Let b_∞ be the pgf of the stationary distribution. By taking expectation in (7), we get

$$b_\infty(u,v)=b_\infty\{a_1(u,v),a_2(u,v)\}b(u,v)\quad\Leftrightarrow\quad b_\infty(u,v)=\prod_{i=1}^{\infty}b\{a_1^{(h)}(u,v),a_2^{(h)}(u,v)\}.$$

The latter expression can be simplified in the special case considered in Corollary 1.

Proposition 4. *If*

(i) the processes $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are mutually independent and Poisson $\mathcal{P}(\lambda_1)$, $\mathcal{P}(\lambda_2)$ distributed; and
(ii) $q_1=q_2=0$ (which is possible only if $\alpha_{11}+\alpha_{21}<1$ and $\alpha_{12}+\alpha_{22}<1$),

then

(i) the stationary distribution of the process $\boldsymbol{X}_t$ is such that $X_{1,t}$ and $X_{2,t}$ are independent, Poisson $\mathcal{P}(\lambda_{1,\infty})$, $\mathcal{P}(\lambda_{2,\infty})$ distributed, with parameters $\lambda_{1,\infty},\lambda_{2,\infty}$ given by

$$(\lambda_{1,\infty},\lambda_{2,\infty})^\top=(Id-A)^{-1}(\lambda_1,\lambda_2)^\top,$$

(ii) and the sum process $Y_t=X_{1,t}+X_{2,t}$ is Poisson INAR(1), with Poisson parameter $\lambda_1+\lambda_2$ and probability parameter $\alpha=1-(\lambda_1+\lambda_2)/(\lambda_{1,\infty}+\lambda_{2,\infty})$.

Proposition 4 is the bivariate analog of the well-known result that the stationary distribution of a univariate Poisson-INAR(1) process is Poisson [34]. It is interesting to compare it with Proposition 2, since in both cases we have assumed $q_1=q_2=0$ and the sum process is INAR(1). On the one hand, Proposition 2 requires condition $\alpha_{11}+\alpha_{21}=\alpha_{12}+\alpha_{22}$, but leaves the distribution of the innovation $(\boldsymbol{\epsilon}_t)$ unconstrained. On the other hand, Proposition 4 does not restrict matrix A, but is based on the independent Poissonian assumption of $(\boldsymbol{\epsilon}_t)$.

2.4.3. The marginal moments

The simple conditional expectation allows us to derive the first two marginal moments of the process. We have the following result.

Proposition 5. *The marginal expectation of the process is given by*

$$\mathrm{E}(\boldsymbol{X}_t)=(Id-A)^{-1}\mathrm{E}(\boldsymbol{\epsilon}_t) \tag{15}$$

and the covariance matrix is

$$\Gamma(0)=\begin{bmatrix}\mathrm{var}(X_{1,t}), & * \\ \mathrm{cov}(X_{1,t},X_{2,t}), & \mathrm{var}(X_{2,t})\end{bmatrix}=\sum_{h=0}^{\infty}A^h\{\gamma_\epsilon+\mathrm{E}(X_{1,t})\gamma_{12}+\mathrm{E}(X_{2,t})\gamma_{34}\}(A^\top)^h, \tag{16}$$

Table 1
Comparison of over-dispersion coefficients and correlation coefficients of the three specifications of BINAR(1) model defined in Eq. (2).

	$\mathrm{var}(X_{1,t})/\mathrm{E}(X_{1,t})$	$\mathrm{var}(X_{1,t})/\mathrm{E}(X_{1,t})$	$\mathrm{corr}(X_{1,t}, X_{2,t})$
Model 1	1.95	1.80	0.84
Model 2	1.47	1.40	0.55
Model 3	1	1	0

whereas the autocovariance matrices are

$$\forall_{h\in\mathbb{N}} \quad \Gamma(h) = \begin{bmatrix} \mathrm{cov}(X_{1,t}, X_{1,t-h}), & \mathrm{cov}(X_{1,t}, X_{2,t-h}) \\ \mathrm{cov}(X_{2,t}, X_{1,t-h}), & \mathrm{cov}(X_{2,t}, X_{2,t-h}) \end{bmatrix} = A^h \Gamma(0). \tag{17}$$

Thus, since the matrix A is nonnegative, the variances $\mathrm{var}(X_{1,t})$, $\mathrm{var}(X_{2,t})$ are increasing in q_1, q_2, ceteris paribus, whereas the marginal expectations do not depend on these two probabilities. Thus the marginal over-dispersion coefficients $\mathrm{var}(X_{1,t})/\mathrm{E}(X_{1,t})$, $\mathrm{var}(X_{2,t})/\mathrm{E}(X_{2,t})$ have a wider range than in the independent BINAR model.

To illustrate this greater flexibility, we compare three BINAR models (2) which differ only by q_1, q_2. We set

$$A = \begin{bmatrix} 0.5, & 0.3 \\ 0.4, & 0.5 \end{bmatrix}$$

and assume $\epsilon_{1,t}$, $\epsilon_{2,t}$ to be independent, $\mathcal{P}(1)$ distributed, whereas q_1, q_2 are specified as follows:

(i) In Model 1, $q_1 = \min(\alpha_{1,1}, \alpha_{1,2}) = 0.4$, $q_2 = \min(\alpha_{2,1}, \alpha_{2,2}) = 0.3$, i.e., both bivariate Bernoulli variables $(Z_1, Z_2)^\top$ and $(Z_3, Z_4)^\top$ have maximal, positive correlation.
(ii) In the [independent BINAR(1)] Model 2, $q_1 = \alpha_{1,1}\alpha_{1,2} = 0.2$, $q_2 = \alpha_{2,1}\alpha_{2,2} = 0.15$, i.e., both bivariate Bernoulli variables have zero correlation.
(iii) In Model 3, $q_1 = q_2 = 0$, i.e., both bivariate Bernoulli variables have minimal, negative correlation.

Table 1 reports the over-dispersion coefficients and correlation coefficients computed using Proposition 4, under the three above models. Note that in Model 3, the three coefficients are equal to 1, 1, 0, respectively. Indeed, by Proposition 4, the marginal stationary distribution is bivariate independent Poisson. Moreover, as expected, these coefficients are largest (resp. smallest) in Model 1 (resp. Model 3). In other words, by letting q_1, q_2 vary, the dependent BINAR model allows for more flexible over-dispersion and correlation patterns than the benchmark Model 2.

3. Higher-order BINAR processes

3.1. BINAR(p) process

Similar as the INAR(p) process introduced by [14], we define the dependent BINAR (p) process as follows:

Definition 2. We say that process $\boldsymbol{X}_t$ is dependent BINAR(p) if it has the representation

$$\boldsymbol{X}_t = \sum_{i=1}^{p} A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t, \tag{18}$$

where given $\mathcal{F}_{t-1} = \{\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots\}$, bivariate count variables $A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i}$ with $i \in \{1, \ldots, p\}$ are mutually independent, and are independent of $\boldsymbol{\epsilon}_t$. Moreover, $A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i}$ is the sum of $X_{1,t-i}$ independent copies of bivariate Bernoulli variable with marginal (resp. joint) probabilities $\alpha_{11,i}$, $\alpha_{21,i}$ (resp. $q_{1,i}$), as well as $X_{2,t-i}$ independent copies of bivariate Bernoulli variable with marginal (resp. joint) probabilities $\alpha_{12,i}$, $\alpha_{22,i}$ (resp. $q_{2,i}$).

Thus compared with the BINAR(1) process, the extended model (18) has a slightly different interpretation since each individual can produce off-springs of both types at the next p periods, and these production outcomes are independent across these periods.

The stationarity condition of the BINAR(p) process is given below.

Proposition 6. *Process (18) is strictly stationary if and only if*

$$\sum_{i=1}^{p} \alpha_{11,i} < 1, \quad \sum_{i=1}^{p} \alpha_{22,i} < 1, \tag{19}$$

$$\text{and } \Big(1 - \sum_{i=1}^{p} \alpha_{11,i}\Big)\Big(1 - \sum_{i=1}^{p} \alpha_{22,i}\Big) > \Big(\sum_{i=1}^{p} \alpha_{21,i}\Big)\Big(\sum_{i=1}^{p} \alpha_{12,i}\Big). \tag{20}$$

or equivalently, if and only if the eigenvalues of $A_1 + \cdots + A_p$ are smaller than 1 in modulus.

The BINAR(p) process has a weak VAR(p) representation since

$$\mathrm{E}(\boldsymbol{X}_{t+1}|\mathcal{F}_t) = \sum_{i=1}^{p} A_i \boldsymbol{X}_{t+1-p} + \mathrm{E}(\boldsymbol{\epsilon}_{t+1}), \tag{21}$$

and its conditional pgf is

$$\mathrm{E}(u^{X_{1,t}} v^{X_{2,t}}|\mathcal{F}_{t-1}) = \exp\left\{\ln b(u,v) + \sum_{i=1}^{p} X_{1,t-i} \ln a_{i,1}(u,v) + \sum_{i=1}^{p} X_{2,t-i} \ln a_{i,2}(u,v)\right\}, \tag{22}$$

where

$$a_{i,1}(u,v) = q_{1,i}uv + (\alpha_{11,i} - q_{1,i})u + (\alpha_{21,i} - q_{1,i})v + (1 + q_{1,i} - \alpha_{11,i} - \alpha_{21,i}),$$
$$a_{i,2}(u,v) = q_{2,i}uv + (\alpha_{12,i} - q_{2,i})u + (\alpha_{22,i} - q_{2,i})v + (1 + q_{2,i} - \alpha_{12,i} - \alpha_{22,i}).$$

Thus the conditional pgf is exponentially affine in $\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p}$, i.e., the process $\boldsymbol{X}_t$ is CaR of order p [CaR(p)]. Similar as the BINAR(1) process, its h-step-ahead conditional pgf is still exponentially affine.

Corollary 2. *Process defined in (18) is such that, for each $h \geq 1$,*

$$\forall_{u,v>0} \quad \mathrm{E}(u^{X_{1,t+h}} v^{X_{2,t+h}}|\mathcal{F}_t) = \exp\left\{B_{h,0}(u,v) + \sum_{i=1}^{p} B_{h,i}^{\top}(u,v)\boldsymbol{X}_{t+1-i}\right\}, \tag{23}$$

where $B_{h,0}$ and $B_{h,i}$ are univariate and bivariate functions, respectively. For $h = 1$, their values are given by Eq. (22), viz.

$$B_{1,0}(u,v) = \ln\{b(u,v)\},$$
$$B_{1,i}(u,v) = \begin{bmatrix} \ln\left\{q_{1,i}uv + (\alpha_{11,i} - q_{1,i})u + (\alpha_{21,i} - q_{1,i})v + (1 + q_{1,i} - \alpha_{11,i} - \alpha_{21,i})\right\} \\ \ln\left\{q_{2,i}uv + (\alpha_{12,i} - q_{2,i})u + (\alpha_{22,i} - q_{2,i})v + (1 + q_{2,i} - \alpha_{12,i} - \alpha_{22,i})\right\} \end{bmatrix},$$

whereas for $h > 1$, we have the following recursions:

$$B_{h+1,0}(u,v) = B_{h,0}(u,v) + \ln b\{B_{h,1}(u,v)\},$$
$$\forall_{i\in\{1,\ldots,p\}} \quad B_{h+1,i}(u,v) = \mathbf{1}_{i\neq p} B_{h,i+1}(u,v) + B_{1,i}\{B_{h,i}(u,v)\}.$$

The proof is a direct consequence of Eq. (22) and is omitted.

Finally, the marginal mean and variance–covariance matrix of the BINAR(p) also have closed form. Their formulas are derived in Appendix A.3.

3.2. BINAR(∞) process

3.2.1. Definition, stationarity, and memory persistence

A natural extension of the BINAR(p) model is to let the order p go to infinity. More precisely, we have the following result.

Definition 3. We say that the process $\boldsymbol{X}_t$ is dependent BINAR(∞) if it has the representation

$$\forall_t \quad \boldsymbol{X}_t = \sum_{i=1}^{\infty} A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t, \tag{24}$$

where the variables $A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i}$ are defined in the same way as in Definition 2.

In the above definition, the partial sum $\sum_{i=1}^{p} A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i}$ given $\mathcal{F}_{t-1}$ converges almost surely to a non degenerate limit when $p \to \infty$. Moreover, since all the terms on the right-hand side are integer-valued, the variables $A_i(q_{1,i}, q_{2,i}) \circ \boldsymbol{X}_{t-i}$ are null for i larger than a certain stochastic threshold τ_t.

To our knowledge, this is the first infinite-order INAR-type model in the literature; see also [26] for infinite-order, univariate INARCH process. Let us first provide its stationarity condition.

Proposition 7. *The process (18) is strictly stationary if and only if*

$$\sum_{i=1}^{\infty} \alpha_{11,i} < 1, \quad \sum_{i=1}^{\infty} \alpha_{22,i} < 1, \tag{25}$$

$$\sum_{i=1}^{\infty} \alpha_{12,i} < \infty, \quad \sum_{i=1}^{\infty} \alpha_{21,i} < \infty, \tag{26}$$

$$\Big(1-\sum_{i=1}^{\infty}\alpha_{11,i}\Big)\Big(1-\sum_{i=1}^{\infty}\alpha_{22,i}\Big) > \Big(\sum_{i=1}^{\infty}\alpha_{21,i}\Big)\Big(\sum_{i=1}^{\infty}\alpha_{12,i}\Big). \tag{27}$$

or equivalently, if and only if matrix $\sum_{i=1}^{\infty} A_i$ is finite and its eigenvalues are all smaller than 1 in modulus.

This proposition nests Propositions 3 and 6. Its proof is provided in Appendix A.4.

This stationarity condition has also implications for the ranges of $q_{1,i}, q_{2,i}$, which have to satisfy the analogue of inequalities (3) and (4), for each i. Since the sequences $(\alpha_{11,i})_i$ and $(\alpha_{21,i})_i$ converge to zero, for large i we have $\alpha_{11,i}+\alpha_{21,i} < 1$. Then the ranges become, for large i,

$$0 \le q_{1,i} \le \min(\alpha_{11,i}, \alpha_{21,i}), \quad 0 \le q_{2,i} \le \min(\alpha_{12,i}, \alpha_{22,i}).$$

Finally, the conditional expectation and conditional pgf of a BINAR(∞) process can be deduced from (21) and (23) by replacing p by infinity. Under the stationary condition, these infinite summation are all finite.

3.2.2. Exact simulation

While simulating a BINAR(p) process is straightforward for finite p, this is no longer the case when $p=\infty$. Let us now derive an exact simulation method for the latter. The basic idea is that the infinite summation in (24) is almost surely finite, and thus it suffices to sample the stochastic threshold $\tau_t \in \mathbb{N}$ such that the infinite summation actually stops at order τ_t, i.e., $A_{\tau_t} \circ \boldsymbol{X}_{t-\tau_t}$ is non null but $A_i \circ \boldsymbol{X}_{t-i} = 0$ is zero for any $i \ge \tau_t + 1$. First we define, for each $i \ge 1$,

$$\delta_i(\mathcal{F}_{t-1}) = (1+q_{1,i}-\alpha_{11,i}-\alpha_{21,i})^{X_{1,t-i}}(1+q_{2,i}-\alpha_{12,i}-\alpha_{22,i})^{X_{2,t-i}},$$

which is equal to $\Pr(A_i \circ \boldsymbol{X}_{t-i} = 0|\mathcal{F}_{t-1})$. We also assume, without loss of generality, the following condition.

Assumption 1. For every integer $i \ge 1$, $\alpha_{11,i}+\alpha_{21,i} < 1$ and $\alpha_{12,i}+\alpha_{22,i} < 1$.

Indeed if the first few terms $q_{i,1}, q_{i,2}$ do not satisfy this condition, we can leave the corresponding variables $A_i \circ \boldsymbol{X}_{t-i}$ out of the infinite sum and simulate them separately. This assumption implies that $\delta_i(\mathcal{F}_{t-1}) > 0$, for any i. Then the conditional CDF of the count variable τ is

$$\forall_{i\in\mathbb{N}} \quad \Pr(\tau_t \le i|\mathcal{F}_{t-1}) = \prod_{j=i+1}^{\infty}\delta_j(\mathcal{F}_{t-1}) = F(i+1|\mathcal{F}_{t-1}).$$

This CDF has the following property.

Lemma 2. *The function $i \mapsto F(i|\mathcal{F}_{t-1})$ is nondecreasing. Its upper limit is $\lim_{i\to\infty} F(i|\mathcal{F}_{t-1}) = 1$, whereas its lower limit $F(0|\mathcal{F}_{t-1})$ is strictly positive.*

Since $F(i+1|\mathcal{F}_{t-1})$ can be easily computed, we can simulate $\boldsymbol{X}_t$ given $\mathcal{F}_{t-1}$ as follows:

1. Draw U from the uniform distribution $\mathcal{U}[0,1]$.
2. Find the unique integer $\tau_t \ge 0$ such that $F(\tau_t - 1\mathcal{F}_{t-1}) < U \le F(\tau_t|\mathcal{F}_{t-1})$, where by convention we set $F(-1|\mathcal{F}_{t-1}) = 0$. In particular, by definition $\boldsymbol{X}_{t-\tau_t}$ cannot be zero since $\delta_{\tau_t}(\mathcal{F}_{t-1}) = F(\tau_t|\mathcal{F}_{t-1})/F(\tau_t-1|\mathcal{F}_{t-1}) > 1$.
3. Sample a certain number of independent copies of $A_{\tau_t} \circ \boldsymbol{X}_{t-\tau_t}$ until we obtain the first non-null observation. This is possible due to Assumption 1.
4. Sample random vectors $A_i \circ \boldsymbol{X}_{t-i}$ for $i \in \{1,\ldots,\tau_t - 1\}$, as well as $\boldsymbol{\epsilon}_t$. Then a sample of $\boldsymbol{X}_t$ is given by $\sum_{i=1}^{\tau_t} A_i \circ \boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t$.

3.2.3. A constrained specification with persistent memory

To avoid the curse of dimensionality, in the application of this paper, we will focus on the following constrained BINAR(∞) specification:

$$\forall i \in \mathbb{N}, \quad \alpha_{11,i} = \alpha_{11}/i^d, \quad \alpha_{21,i} = \alpha_{21}/i^d, \quad \alpha_{12,i} = \alpha_{12}/i^d, \quad \alpha_{22,i} = \alpha_{22}/i^d, \quad q_{1,i} = q_1/i^d, \quad q_{2,i} = q_2/i^d, \tag{28}$$

where the power index $d > 1$ to ensure that $\sum_i A_i$ is finite, and for each i, probabilities $q_{1,i}$ and $q_{2,i}$ satisfy the constraint:

$$\max(\alpha_{11,i}+\alpha_{21,i}-1, 0) \le q_{1,i} \le \min(\alpha_{11,i}, \alpha_{21,i}),$$
$$\max(\alpha_{12,i}+\alpha_{22,i}-1, 0) \le q_{2,i} \le \min(\alpha_{12,i}, \alpha_{22,i}).$$

It is easily checked that this is true if and only if these two inequalities hold for $i = 1$.

While in a BINAR(p) model, the autocovariance function decays geometrically in the lag h, a distinct feature of the BINAR(∞) model is that it allows the autocovariance to have a hyperbolic decay rate. More precisely we have the following result.

Lemma 3. *In model (28), the autocovariance matrix $\Gamma(h) = \mathrm{E}[\{\boldsymbol{X}_t - \mathrm{E}(\boldsymbol{X}_t)\}\{\boldsymbol{X}_{t-h} - \mathrm{E}(\boldsymbol{X}_t)\}^\top]$ of the process $\boldsymbol{X}_t$ decays also at the hyperbolic rate d.*

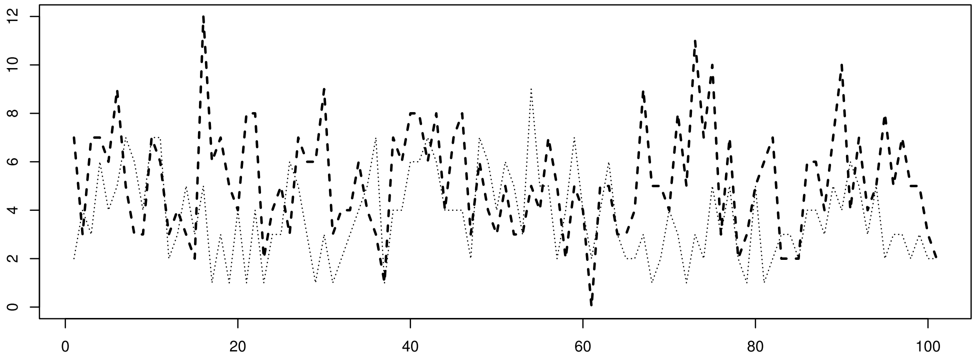

Fig. 1. Simulated trajectory of the BINAR(∞) process defined in (24) under constraint (28). The trajectory of $(X_{1,t})$ is dotted line and that of $(X_{2,t})$ in dashed line.

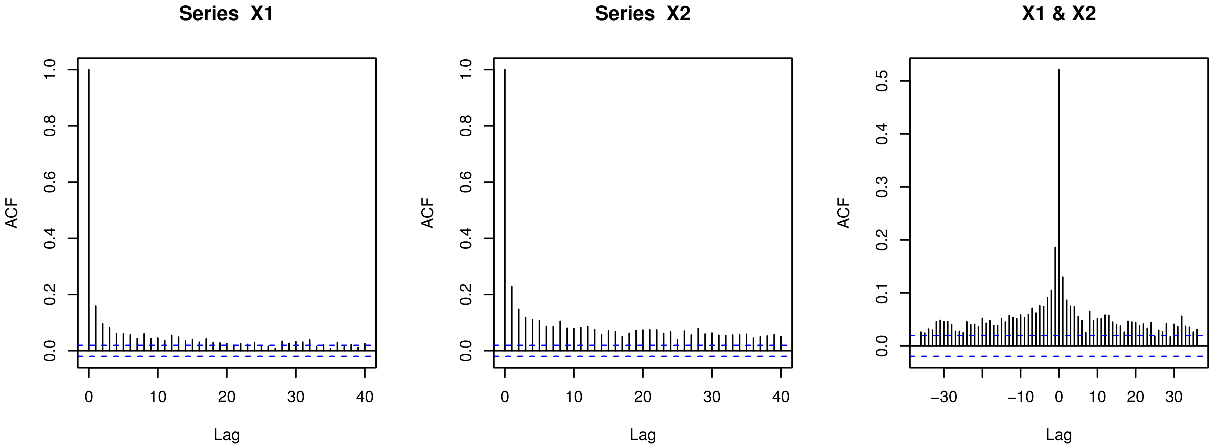

Fig. 2. ACF and CCF function of the same simulated BINAR(∞) process as in Fig. 1. As is expected, both functions decay rather slowly.

This model has a similar spirit as the (univariate) ARCH(∞) model for asset returns [18,43]. Similar as in this latter literature, since $d > 1$, the autocovariance matrix $\Gamma(h)$ is summable, ruling out the possibility that $\sum_{h=0}^{\infty} \Gamma(h) = \infty$.

We plot in Fig. 1 the simulated path of a BINAR(∞) process with

$$A = \begin{bmatrix} 0.12, & 0.03 \\ 0.06, & 0.15 \end{bmatrix}, \quad d = 1.3, \quad q_1 = 0.015, \quad q_2 = 0.03, \quad (\epsilon_{1,t}, \epsilon_{2,t}) \sim \mathcal{BP}(1, 1, 0.5).$$

In this model, the eigenvalues of $(\sum_{i=1}^{\infty} i^{-d})A_1$ are 0.7 and 0.35, respectively; thus the persistence of the process is quite strong.

The simulated means and variances of the two processes are $\hat{\mathrm{E}}(X_{1,t}) = 3.4$, $\hat{\mathrm{E}}(X_{2,t}) = 4.6$, $\hat{\mathrm{var}}(X_{1,t}) = 3.7$, $\hat{\mathrm{var}}(X_{2,t}) = 5.1$, corresponding to a process with mild over-dispersion. In Fig. 2 we also report the autocorrelation functions (ACF) as well as the cross-correlation function (CCF) of the two component processes. These functions are computed using a simulated sample of 10,000 observations.

4. Predictive distributions

It has been argued by McCabe and Martin [32] that one of the essential properties of a count process model is the tractability of the predictive pmf of $\boldsymbol{X}_{t+h}|\mathcal{F}_t$, for both estimation and forecasting purposes.

Indeed, in terms of estimation, there are two natural approaches for the BINAR(p) model. The first one is the Generalized Method of Moments (GMM), based on moment restrictions derived from the conditional pgf [21]. While this approach is computationally simple, it usually induces an efficiency loss. The second approach is the maximum likelihood estimation. Although more efficient, its difficulty lies in the computation of the conditional pmf $\ell(\boldsymbol{X}_t|\mathcal{F}_{t-1})$. Indeed, since this distribution is the convolution of $\sum_{i=1}^{p}(X_{1,t-i}+X_{2,t-i})+1$ bivariate discrete distributions, its expression is highly cumbersome if we apply brute-force convolution; see, e.g., [31] for the expression of $\sum_{i=1}^{n}(Z_{1,j}, Z_{2,j})^{\top}$. This has been identified by [36] as the major downside of higher-order (B)INAR models.

As for forecasting, the conditional expectations of count processes are non integer-valued and thus incompatible with count data. While some approximation methods have been proposed in the univariate INAR framework [24,33], they are generically time consuming and induce significant errors; see [30] for such an analysis. In this section we clarify that in the dependent (and, a fortiori, independent) BINAR(p) model, the conditional pmf $\ell(\boldsymbol{X}_{t+h-1}|\mathcal{F}_{t-1})$ can be computed exactly at any horizon h. This makes the likelihood-based estimation feasible, and eliminates the approximation error induced for the forecasting. For expository purposes, we focus on BINAR(p) models with a finite, potentially large p. We first consider the horizon $h=1$, before explaining how to adapt the algorithm to higher horizons.

4.1. One-step-ahead predictive distribution

Our aim here is to compute the probabilities $\Pr\{\boldsymbol{X}_t = (x_1,x_2)^\top|\mathcal{F}_{t-1}\}$ simultaneously for all couples $(x_1,x_2) \in [|0,m|]\times[|0,n|]$, where the bounds m,n are chosen such that the residual probability $\Pr\{X_{1,t} > m \text{ or } X_{2,t} > n|\mathcal{F}_{t-1}\}$ is negligible.

Let us first remark that for any count process, the conditional pgf and pmf are linked via

$$\forall_{u,v\geq 0}\quad \mathrm{E}(u^{X_{1,t}}v^{X_{2,t}}|\mathcal{F}_{t-1}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\Pr\{\boldsymbol{X}_t = (i,j)^\top|\mathcal{F}_{t-1}\}u^i v^j.$$

Thus $\Pr\{\boldsymbol{X}_t = (x_1,x_2)^\top|\mathcal{F}_{t-1}\}$ is equal to the (x_1,x_2)th order coefficient in the above Taylor expansion with respect to (u,v), at $(0,0)$. Let us now make use of the simple conditional pgf to compute the Taylor expansion up to order (m,n). First, we rewrite (22) into

$$\begin{aligned}\mathrm{E}(u^{X_{1,t}}v^{X_{2,t}}|\mathcal{F}_{t-1}) &= \exp\left\{\ln b(0,0) + \sum_{i=1}^{p}X_{1,t-i}\ln a_1(0,0) + \sum_{i=1}^{p}X_{2,t-i}\ln a_2(0,0)\right\}\\ &\quad\times\exp\left\{\ln\frac{b(u,v)}{b(0,0)} + \sum_{i=1}^{p}X_{1,t-i}\ln\frac{a_{i,1}(u,v)}{a_{i,1}(0,0)} + \sum_{i=1}^{p}X_{2,t-i}\ln\frac{a_{i,2}(u,v)}{a_{i,2}(0,0)}\right\}.\end{aligned}$$

Then we perform the (m,n)th order Taylor expansion of $\ln\{b(u,v)/b(0,0)\}$ and $\ln\{a_{i,j}(u,v)/a_{i,j}(0,0)\}$ at $(0,0)$ for each $i\in\{1,\ldots,p\}$ and $j\in\{1,2\}$. For most standard bivariate count distributions, $\ln b(u,v)$ is Taylor-expandable. Examples include the bivariate Poisson distribution, see Eq. (10), the bivariate Hermite distribution, see Eq. (11), as well as the bivariate negative binomial distribution [15], with pgf

$$\forall_{u,v\geq 0 \text{ such that } b_1u+b_2v<1}\quad b(u,v) = (1-b_1-b_2)^\theta/(1-b_1u-b_2v)^\theta,$$

where $b_1,b_2,\theta>0$, $1-b_1-b_2>0$.

As for $\ln\{a_{i,j}(u,v)/a_{i,j}(0,0)\}$, we have

$$\ln a_{i,j}(u,v)/a_{i,j}(0,0) = \sum_{k=1}^{\infty}(-1)^{k-1}P_{i,j}^k(u,v)/k = \sum_{k=1}^{m+n}(-1)^{k-1}P_{i,j}^k(u,v)/k + o_{m,n}(u,v), \tag{29}$$

where, for all $i\in\{1,\ldots,p\}$,

$$\begin{aligned}P_{i,1}(u,v) &= \frac{a_{i,1}(u,v)}{a_{i,1}(0,0)} - 1 = \frac{q_{1,i}uv + (\alpha_{11,i}-q_{1,i})u + (\alpha_{21,i}-q_{1,i})v}{1+q_{1,i}-\alpha_{11,i}-\alpha_{21,i}},\\ P_{i,2}(u,v) &= \frac{a_{i,2}(u,v)}{a_{i,2}(0,0)} - 1 = \frac{q_{2,i}uv + (\alpha_{12,i}-q_{2,i})u + (\alpha_{22,i}-q_{2,i})v}{1+q_{2,i}-\alpha_{12,i}-\alpha_{22,i}},\end{aligned}$$

are polynomials in u and v without constant term, whereas $o_{m,n}(u,v)$ represents the omitted higher-order terms in the expansion. Let us explain why the truncation in (29) stops at order $m+n$. The polynomial $P_{i,j}^k(u,v)$ is a linear combination of terms $u^{k_1}v^{k_2}$, where $k_1+k_2\geq k$. Thus if $k>m+n$, then either $k_1>m$ or $k_2>n$, and $u^{k_1}v^{k_2}$ is omitted in the (m,n)th order Taylor expansion. Therefore, we only need to expand recursively each $P_{i,j}^k(u,v)$ for $k\in\{1,\ldots,m+n\}$, $i\in\{1,\ldots,p\}$ and $j\in\{1,2\}$ and truncate these polynomials at order (m,n). This latter can be achieved by the following algorithm.

Proposition 8. *If we represent the (m,n)th order truncation of a polynomial*

$$P(u,v) = \sum_{k_1=0}^{m}\sum_{k_2=0}^{n}c_{k_1,k_2}u^{k_1}v^{k_2} + o_{m,n}(u,v),$$

by the column vector

$$(\underbrace{c_{0,0},c_{0,1},\ldots,c_{0,n}}_{n+1\text{ terms}},\underbrace{c_{1,0},c_{1,1},\ldots,c_{1,n}}_{n+1\text{ terms}},\ldots,\underbrace{c_{m,0},c_{m,1},\ldots,c_{m,n}}_{n+1\text{ terms}})^\top\in\mathbb{R}^{(m+1)(n+1)},$$

then the (m, n)th order truncation of polynomial $P^k(u, v)$ is represented by the column vector

$$\underbrace{\begin{bmatrix} M_0 & 0 & 0 & 0 & \cdots \\ M_1 & M_0 & 0 & \cdots & \cdots \\ M_2 & M_1 & M_0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ M_m & M_{m-1} & \cdots & \cdots & M_0 \end{bmatrix}^k}_{=M} \begin{bmatrix} e_1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

where column vector $e_1 = (1, 0, 0, \ldots, 0)^\top \in \mathbb{R}^{n+1}$; square matrix $M \in \mathcal{M}_{(m+1)(n+1)}(\mathbb{R})$ is $(m + 1) \times (m + 1)$ block lower triangular Toeplitz, i.e.,

$$M = \{\mathbf{1}(i \geq j)M_{i-j}\}_{0\leq i,j\leq m},$$

and the block matrices $M_i \in \mathcal{M}_{n+1}(\mathbb{R})$ are themselves lower triangular Toeplitz, viz.

$$\forall_{i\in\{0,\ldots,m\}} \quad M_i = \{\mathbf{1}(k_1 \geq k_2)c_{i,k_1-k_2}\}_{0\leq k_1,k_2\leq n}.$$

As an illustration, if $m = n = 2$ we have

$$M = \left[\begin{array}{ccc|ccc|ccc} c_{00} & 0 & 0 & & & & & & \\ c_{01} & c_{00} & 0 & & \mathbf{0} & & & \mathbf{0} & \\ c_{02} & c_{01} & c_{00} & & & & & & \\ \hline c_{10} & 0 & 0 & c_{00} & 0 & 0 & & & \\ c_{11} & c_{10} & 0 & c_{01} & c_{00} & 0 & & \mathbf{0} & \\ c_{12} & c_{11} & c_{10} & c_{02} & c_{01} & c_{00} & & & \\ \hline c_{20} & 0 & 0 & c_{10} & 0 & 0 & c_{00} & 0 & 0 \\ c_{21} & c_{20} & 0 & c_{11} & c_{10} & 0 & c_{01} & c_{00} & 0 \\ c_{22} & c_{21} & c_{20} & c_{12} & c_{11} & c_{10} & c_{02} & c_{01} & c_{00} \end{array}\right].$$

This matrix has $(m + 1)^2 = 9$ blocks, each block is a $(n + 1) \times (n + 1) = 3 \times 3$ matrix.

The proof of this proposition is obvious and omitted.

After Taylor-expanding $\ln b(u, v)$ and $\ln\{1 + P_{i,j}(u, v)\}$ for $i \in \{1, \ldots, p\}$ and $j \in \{1, 2\}$, we sum up these expansions and get

$$\begin{aligned} \mathrm{E}(u^{X_{1,t}} v^{X_{2,t}} | \mathcal{F}_{t-1}) &= c \exp\left\{ \sum_{k_1=0}^{m} \sum_{k_2=0}^{n} f_{k_1,k_2}(\mathcal{F}_{t-1}) u^{k_1} v^{k_2} + R_m(u, v) \right\} \\ &= c \sum_{k=0}^{m+n} \frac{1}{k!} \left\{ \sum_{k_1=0}^{m} \sum_{k_2=0}^{n} f_{k_1,k_2}(\mathcal{F}_{t-1}) u^{k_1} v^{k_2} \right\}^k + o_{m,n}(u, v), \end{aligned} \tag{30}$$

where the coefficients $f_{k_1,k_2}(\mathcal{F}_{t-1})$ are affine in $\mathcal{F}_{t-1}$, whereas

$$c = \exp\left\{ \sum_{i=1}^{p} X_{1,t-i} \ln(1 + q_{1,i} - \alpha_{11,i} - \alpha_{21,i}) + \sum_{i=1}^{p} X_{2,t-i} \ln(1 + q_{2,i} - \alpha_{12,i} - \alpha_{22,i}) + \ln b(0, 0) \right\}.$$

In Eq. (30), the expansion stops at order $m+n$ for the same reason as in (29). Then we apply Proposition 8 and obtain the (m, n)th order truncation of polynomials

$$\left\{ \sum_{k_1=0}^{m} \sum_{k_2=0}^{n} f_{k_1,k_2}(\mathcal{F}_{t-1}) u^{k_1} v^{k_2} \right\}^k$$

for each $k \in \{1, \ldots, m + n\}$. Finally by coefficient matching we get the pmf of $\boldsymbol{X}_t | \mathcal{F}_{t-1}$.

In terms of computational cost, the Taylor expansions of $\ln b(u, v)$ and $\ln\{1+P_{i,j}(u, v)\}$ with $i \in \{1, \ldots, p\}$ and $j \in \{1, 2\}$ are conducted only once when t varies. Thus for each t, the computation of $\ell(\boldsymbol{X}_t | \mathcal{F}_{t-1})$ involves essentially the computation of the right-hand side of Eq. (30), whose cost is independent of p. Thus this method is applicable even for large p.

The tractability of the conditional distribution has several important implications. First, it allows for efficient maximum likelihood estimation. In the Online Supplement, we propose a small comparison between the MLE and a GMM estimator via Monte Carlo experiments. Second, the likelihood function can also be used for model selection via the information

Table 2
CPU time of the two forecasting methods applied to BINAR(1) and BINAR(4) models. For the exact approach, we compute the value of the conditional pmf $\ell((i,j)^\top|\mathcal{F}_T)$ for any i, j ranging from 0 to 15, although in the table we only report their values for i, j non larger than 8.

$p = 1$, $\boldsymbol{X}_T = (1,4)^\top$			
Method	Exact method	Simulation-based method	Simulation-based method
Number of draws	0	$N = 10{,}000$	$N = 100{,}000$
CPU time	0.01 s	0.05 s	0.5 s
$p = 4$, $\boldsymbol{X}_T = \boldsymbol{X}_{t-1} = \boldsymbol{X}_{T-2} = \boldsymbol{X}_{T-3} = (1,4)^\top$			
Method	Exact method	Simulation-based method	Simulation-based method
Number of draws	0	$N = 10{,}000$	$N = 100{,}000$
CPU time	0.01 s	0.2 s	2 s

criteria [41], as an alternative to the Box–Jenkins approach [7]. Third, it allows to conduct likelihood ratio type tests for statistical significance of the parameters.

4.2. Comparison with the simulation-based forecasting approach

Let us now illustrate how the exact forecasting approach fares against the state-of-the-art simulation-based method [24,32] when it comes to the computation of the one-step-ahead predictive distribution. While this method is general and applies to other non-BINAR models, it will be shown that one of the advantages of BINAR models is that the exact approach outperforms significantly the simulation approach, both in terms of computational time and forecasting accuracy. Note that McCabe et al. [33] proposed another method to approximate the predictive pmf. They approximated the univariate INAR(p) process by a Markov chain with S states, where S is a large integer. This method involves the computation of matrices of dimension $S^{2p} \times S^{2p}$, which is extremely cumbersome when $p \geq 2$.

To this end we consider the BINAR(p) model with

$$A_i = \frac{1}{i^d}\begin{bmatrix} 0.12, & 0.06 \\ 0.03, & 0.15 \end{bmatrix}, \quad i \in \{1,\ldots,p\}, \quad q_1 = 0.015, \quad q_2 = 0.03, \quad \boldsymbol{\epsilon}_t \sim \mathcal{BP}(2,2,2)$$

Given the past observation $\mathcal{F}_T$, the simulation-based method consists in drawing a large number of possible future values $\boldsymbol{X}_{T+1}^{(n)}$ with $n \in \{1,\ldots,N\}$. Then the conditional pmf is approximated by

$$\forall_{i,j\in\mathbb{N}} \quad \Pr\{\boldsymbol{X}_{T+1} = (i,j)^\top|\mathcal{F}_T\} \approx \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\boldsymbol{X}_{T+1}^{(n)} = (i,j)^\top\}.$$

We first report in Table 2 the run time of the two methods (for the simulation-based method we consider two values for the number of draws). Both methods are implemented in R using the same laptop (intel i5, 3.0 GHz, 8 GB RAM) and the program is available from the authors upon request.

We see that the exact method is 10 to 100 times faster than the simulation method. Moreover this ratio becomes larger when the order p increases. This is expected since the run time of the simulation-based method is roughly proportional to p, whereas as we see in the previous section, it does not depend on p in our approach.

Let us now evaluate the approximation error of the simulation-based approach. We focus on the above BINAR(1) model, and express, in Table 3, the approximate conditional pmf's as a percentage of the corresponding exact values.

We can see that the approximation error of the simulation approach is substantial, even with a huge number of draws ($N = 100{,}000$). This is particularly the case for the probabilities of "extreme events", i.e., when either $X_{1,t+1}$ or/and $X_{2,t+1}$ is large. This is a serious downside of the standard simulation approach, since in finance, predicting extreme events is key to the risk management.

4.3. Multiple-step-ahead predictive distributions

Let us now adapt the above algorithm for the computation of $\ell(\boldsymbol{X}_{t+h}|\mathcal{F}_t)$. For expository purposes, we focus on the BINAR(1) process. To Taylor-expand the conditional pgf given in Proposition 1, we use the following procedure:

1. First, we use Proposition 8 to compute the (m,m)th-order Taylor expansion of $\ln a_1^{(h)}(u,v)$, $\ln a_2^{(h)}(u,v)$ and $b^{(h)}(u,v)$ at $(u,v) = (0,0)$, where m is chosen such that the probability of either component processes taking values larger than m is negligible. Note that although by Proposition 1, $a_1^{(h)}$ and $a_2^{(h)}$ are 2^hth-order polynomial in (u,v), which can be large for large h, only the terms of degree lower than $\min(m, 2^h)$ contribute to the (m,m)th Taylor expansion. Thus this step involves roughly the same computational effort for different values of h.

Table 3
Conditional pmf (upper panel) as well as the relative accuracy of the simulation approach compared to the exact approach, for $N = 10{,}000$ draws (middle panel) and $N = 100{,}000$ draws (lower panel).

Conditional pmf computed using the exact method					
	$X_{1,T+1}=0$	$X_{1,T+1}=2$	$X_{1,T+1}=4$	$X_{1,T+1}=6$	$X_{1,T+1}=8$
$X_{2,T+1}=0$	0.00096	0.00248	0.00104	0.00017	0.00001
$X_{2,T+1}=2$	0.00324	0.02270	0.01944	0.00552	0.00074
$X_{2,T+1}=4$	0.00172	0.02491	0.04171	0.02079	0.00447
$X_{2,T+1}=6$	0.00035	0.00885	0.02625	0.02250	0.00782
$X_{2,T+1}=8$	0.00003	0.00145	0.00698	0.00975	0.00542
Relative accuracy of the simulation-based method, $N = 10,000$					
	$X_{1,T+1}=0$	$X_{1,T+1}=2$	$X_{1,T+1}=4$	$X_{1,T+1}=6$	$X_{1,T+1}=8$
$X_{2,T+1}=0$	72%	132%	105%	115%	0%
$X_{2,T+1}=2$	111%	101%	104%	104%	93%
$X_{2,T+1}=4$	87%	96%	98%	99%	109%
$X_{2,T+1}=6$	85%	103%	106%	108%	91%
$X_{2,T+1}=8$	0%	89%	91%	108%	101%
Relative accuracy of the simulation-based method, $N = 100,000$					
	$X_{1,T+1}=0$	$X_{1,T+1}=2$	$X_{1,T+1}=4$	$X_{1,T+1}=6$	$X_{1,T+1}=8$
$X_{2,T+1}=0$	118%	99%	105%	92%	65%
$X_{2,T+1}=2$	101%	97%	101%	96%	115%
$X_{2,T+1}=4$	96%	98%	100%	100%	100%
$X_{2,T+1}=6$	99%	96%	101%	97%	98%
$X_{2,T+1}=8$	80%	95%	97%	105%	100%

2. Then we compute the (m, m)th order Taylor expansion of

$$\begin{aligned}
&\exp\{X_{1,t}\ln a_1^{(h)}(u,v) + X_{2,t}\ln a_2^{(h)}(u,v) + \ln b^{(h)}(u,v)\} \\
&\qquad = c\exp\left\{X_{1,t}\ln\frac{a_1^{(h)}(u,v)}{a_1^{(h)}(0,0)} + X_{2,t}\ln\frac{a_2^{(h)}(u,v)}{a_2^{(h)}(0,0)} + \ln b^{(h)}(u,v)\right\} \\
&\qquad = c\sum_{k=0}^{2m}\frac{1}{k!}\left\{X_{1,t}\ln\frac{a_1^{(h)}(u,v)}{a_1^{(h)}(0,0)} + X_{2,t}\ln\frac{a_2^{(h)}(u,v)}{a_2^{(h)}(0,0)} + \ln\frac{b^{(h)}(u,v)}{b(0,0)}\right\}^k + o_{m,m}(u,v),
\end{aligned}$$

where $c = \exp\{X_{1,t}\ln a_1^{(h)}(0,0) + X_{2,t}\ln a_2^{(h)}(0,0) + \ln b(0,0)\}$. This is conducted using the same method as for $h = 1$. Finally we deduce $\Pr\{\boldsymbol{X}_{t+h} = (i,j)^\top | \mathcal{F}_t\}$ for any $(i,j) \in [|0,m|]^2$ by coefficient matching.

5. An application to mutual fund flows

5.1. The mutual fund industry

Mutual funds (MF) are investment vehicles who invest in a wide range of assets ranging from liquid ones such as stocks, bonds, to highly illiquid ones such as hedge funds. Their clients include, for instance, insurance companies, private banks, large corporations as well as retail investors (In some funds, including the one we study, retail investors' orders are first centralized by a broker before being transferred to the fund. Thus from the fund's point of view, its client is the broker). They are traditionally much less regulated than commercial banks, but this potential loophole has recently received much attention from the regulators.

Most MF are open-ended, i.e., they allow investors to purchase new shares, or redeem their shares on a daily basis. Thus the size of the MF can feature important short-term fluctuations, making them vulnerable to liquidity risk. In particular, during a market turmoil, investors' redemption decisions tend to cluster. If the fund manager's cash holding is insufficient to meet the redemption requests, he/she might be forced to sell its illiquid assets, whose market liquidity would have also plunged due to the crisis. Such fire selling usually leads to significant investment loss, which in turn creates panics and triggers further redemptions. This phenomenon is called fund run.

In contrast, while inflow, i.e., the purchase orders of the MF, or cash holding can offset the outflow due to redemption, a sudden large inflow also dilutes the investment performance of the fund due to the lack of immediate investment opportunities. Thus they can also trigger subsequent (large) outflows. As a consequence, it is essential for the fund manager to monitor in parallel the outflow and the inflow pattern of its clients, on a daily basis. The current MF literature usually focuses on the outflow only, or the net outflow, i.e., the difference between the outflow volume and inflow volume [12,38]. Moreover, many of these studies are based on weekly data only.

Finally, while prior studies focus on the volume of the outflow/inflow, our attention is on the number of purchases $X_{1,t}$ and redemption orders $X_{2,t}$. These variables are closely related since on each trading day, the volume $Y_{1,t}$ and $Y_{2,t}$ of

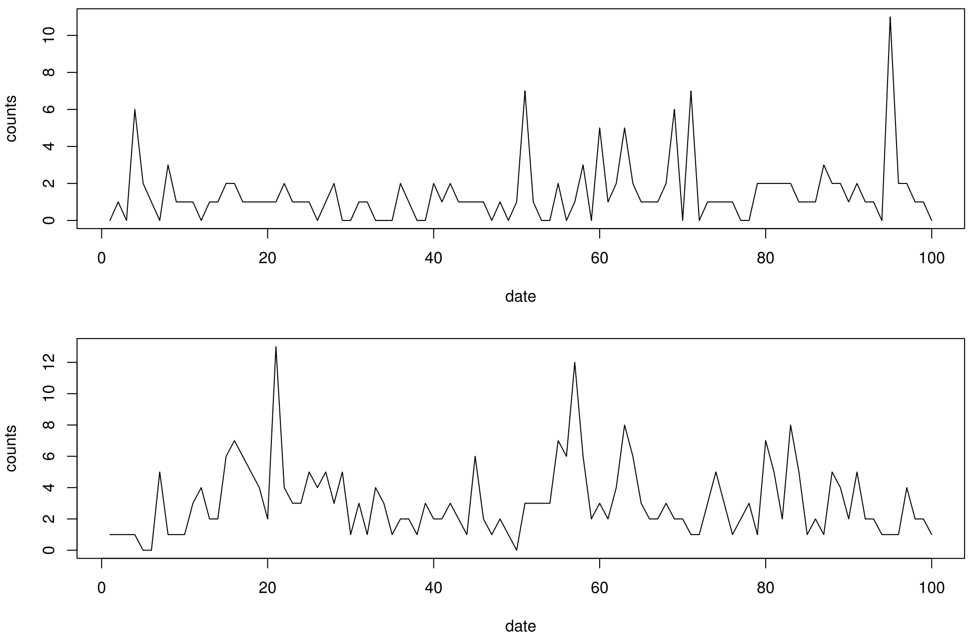

Fig. 3. Joint trajectory of the purchase count (upper panel) and redemption count (lower panel) during the same 100 trading days.

the outflow/inflow have the compound representation

$$Y_{1,t} = \sum_{j=1}^{X_{2,t}} S_{j,t}, \quad Y_{2,t} = \sum_{j=1}^{X_{2,t}} B_{j,t}, \tag{31}$$

where $S_{j,t}$ (resp. $B_{j,t}$) denotes the volume of the jth redemption (resp. purchase). Moreover, when j and t vary, $(S_{j,t})$ and $(B_{j,t})$ can be reasonably assumed to be i.i.d.

In this framework, our preference for studying the counts rather than the volume is motivated by the following reasons. First, very often a large outflow volume is due to the redemption order of a large investor. These "VIP" clients usually have privileged relationship with the fund manager and in the case of a large redemption, they also tend to (privately) inform the fund manager sufficiently in advance so that the latter can avoid a massive fire selling. However, a large $X_{2,t}$ spells a collective withdrawal, i.e., the fund run, which is the most dangerous scenario. More importantly, the non-linear forecasting of the net outflow, i.e., $Y_{2,t} - Y_{1,t}$, can be easily deduced from representation (31). Indeed, the conditional Laplace transform of $Y_{2,t} - Y_{1,t}$ given $\mathcal{F}_{t-1}$ is

$$\forall_{u\in\mathbb{R}} \quad \mathrm{E}\{e^{-u(Y_{2,t}-Y_{1,t})}|\mathcal{F}_{t-1}\} = \mathrm{E}\Big[\{\mathrm{E}(e^{uB_{1,t}})\}^{X_{2,t}}\{\mathrm{E}(e^{-uS_{1,t}})\}^{X_{1,t}}|\mathcal{F}_{t-1}\Big]. \tag{32}$$

The right-hand side is the conditional pgf of the count process evaluated at $\{\mathrm{E}(e^{uS_{1,t}}), \mathrm{E}(e^{-uB_{1,t}})\}$, which has closed form under suitable distributional assumptions on $S_{1,t}$ and $B_{1,t}$ (such as gamma). Thus the left-hand side of (32) is readily available. Then the conditional Value-at-Risk of the net outflow, say, can be accurately approximated without simulation [19]. As a consequence, in the paper we will only focus on counts.

5.2. Data description

Our dataset comes from a French equity-focused MF. We observe the daily number of purchase orders $X_{1,t}$ and redemption orders $X_{2,t}$, during some 1000 trading days. Figs. 3–5 plot the trajectory of the two count processes during a sub-sample of 100 trading days, as well as the histograms of $X_{1,t}$ and $X_{2,t}$.

We report below the empirical marginal moments of the two processes.

Both processes feature mild unconditional over-dispersion, a typical feature of BINAR processes. Fig. 5 plots their autocorrelation patterns.

These ACF's and CCF decay rather slowly, resembling the simulated patterns of Section 3.2.3 for a model with hyperbolically decaying coefficients A_i, $q_{1,i}$, $q_{2,i}$. Thus we will focus on the estimation of this latter model. Note that we can interpret the thinning part of model (2) as trades by clients who also made purchase/sell orders in the previous periods. Then the binomial distributional assumption can be explained by the fact that usually clients only trade on a daily basis. Even if they place several orders within a day, usually the fund only executes the orders once each day at the market

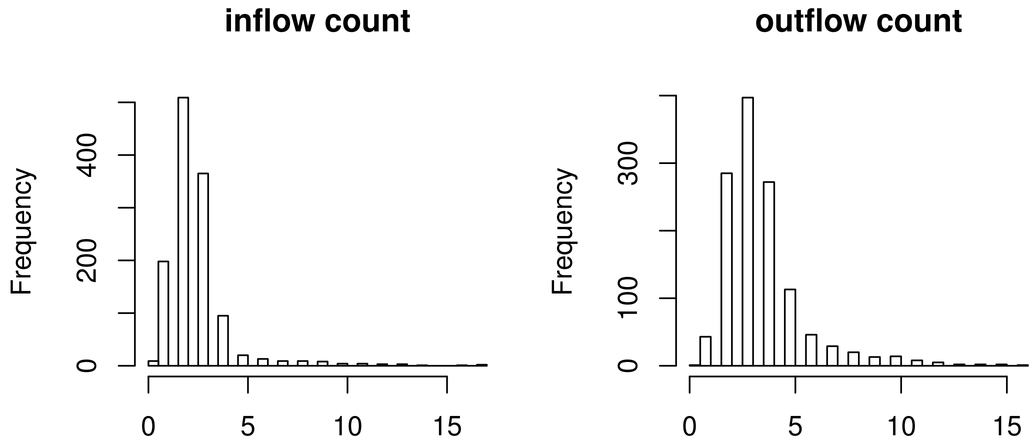

Fig. 4. Histogram of $(X_{1,t})$ (left panel) and $(X_{2,t})$ (right panel) counts. We can see that the probability of $X_{1,t}$ and $X_{2,t}$ taking large values is extremely small.

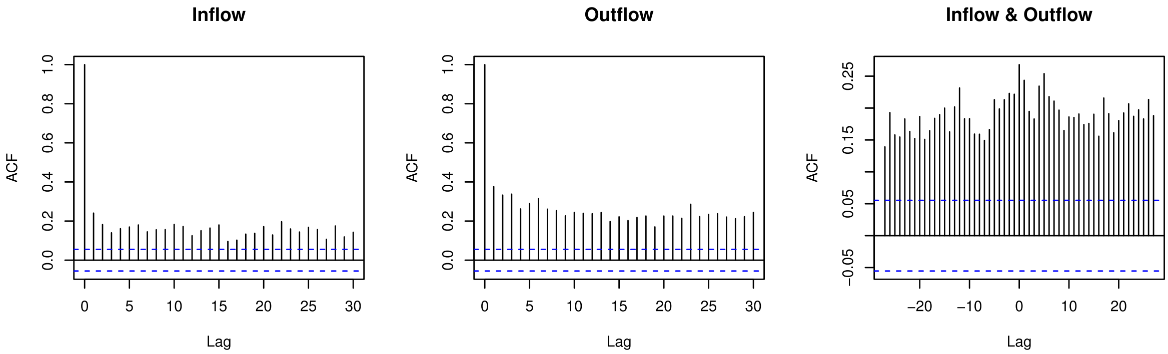

Fig. 5. ACF of inflow count process $(X_{1,t})$ (left panel) and outflow process count $(X_{2,t})$ (middle panel), as well as the CCF (right panel).

Table 4
Empirical (marginal) moments of the two count processes $(X_{1,t})$ and $(X_{2,t})$.

$\hat{\mathrm{E}}(X_{1,t})$	$\hat{\mathrm{E}}(X_{2,t})$	$\hat{\mathrm{var}}(X_{1,t})$	$\hat{\mathrm{var}}(X_{2,t})$	$\hat{\mathrm{corr}}(X_{1,t}, X_{2,t})$
2.63	3.06	3.67	3.97	0.27

closure, and thus only one aggregate order is counted for each client. Thus this interpretation suggests model (2) with the further constraint that $q_1 = q_2 = 0$, i.e., clients do not make simultaneous buy and sell orders as these two will partially cancel out. In practice, in the application, although this constrained version of model (2) is easier to interpret, we have considered a BINAR(∞) as this latter reflects better the persistence of the ACF's.

Since the CCF indicates a positive (marginal) correlation between the two processes, it is reasonable to assume that the innovation process $\boldsymbol{\epsilon}_t$ features also positive correlation between its two components. Moreover, Table 4 suggests that the degree of unconditional over-dispersion of the two count variables is rather weak, making the assumption of Poisson marginal distribution plausible for $\boldsymbol{\epsilon}_t$. Thus we assume the distribution of the latter to be bivariate Poisson $\mathcal{BP}(\lambda_1, \lambda_2, \lambda_3)$. Then the set of parameters is

$$\theta = (\alpha_{11}, \alpha_{22}, \alpha_{12}, \alpha_{21}, q_1, q_2, \lambda_1, \lambda_2, \lambda_3, d).$$

Let us now interpret the regression parameters $\alpha_{i,j}$. Parameter α_{11} measures how purchase decisions of investors are (positively) correlated, due to the so-called reputation effect. Moreover, since here the investors' benchmark is past yearly performance and our observations concern daily inflow/outflow movements, it is not surprising that this reputation effect decays rather slowly when daily data are used, as is shown by the ACF of $X_{1,t}$. Similarly, parameter α_{22} measures the panic effect among investors. Parameter α_{12} captures the propensity of redemption following large recent inflows, as such inflows might dilute the fund's performance due to lack of sufficient investment opportunities. This is consistent with the CCF given in Fig. 5, which indicates a positive cross correlation between $X_{2,t}$ and the lagged values of $X_{1,t}$. Finally, parameter α_{21} captures the propensity of purchase following large outflows. This can be interpreted as the fund manager's capability of attracting new investment in order to stabilize the fund size.

Table 5
Parameter estimates of the constrained BINAR(∞) model along with the corresponding standard errors in parentheses, obtained by numerically inverting the Hessian matrix.

Parameter	Estimate	Parameter	Estimate
α_{11}	0.174 (0.055)	α_{21}	0.036 (0.016)
α_{12}	0.120 (0.052)	α_{22}	0.285 (0.121)
q_1	0.022 (0.012)	q_2	0.017 (0.007)
λ_1	0.712 (0.210)	λ_2	0.695 (0.160)
λ_3	0.354 (0.184)	d	1.830 (0.152)

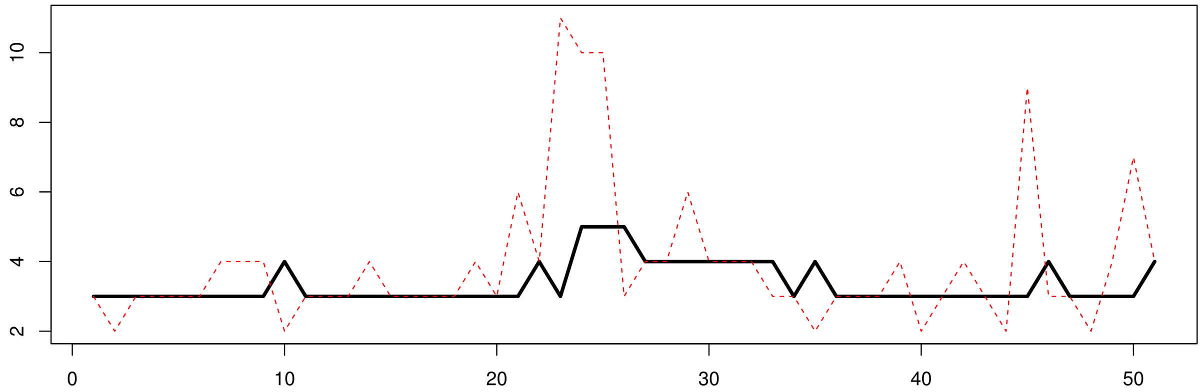

Fig. 6. Joint evolution of $\hat{X}_{2,t+1}$ (in thick full line) and the realized redemption count $X_{2,t+1}$ (in dashed line) during 50 trading days.

5.3. Model estimation

We estimate the parameters by maximum likelihood. A practical difficulty of applying an infinite order model is that our number of observations (approximately 1000) is finite. Thus the "∞" in $\sum_{i=1}^{\infty} A_i \circ \boldsymbol{X}_{t-i}$ should be replaced by a finite, but large p. In the estimation we take $p = 300$, and regard the first p values of X_t as initial values rather than observations. Table 5 reports the maximum likelihood estimates.

The parameter estimate $\hat{\alpha}_{22}$ $(= 0.285)$ is larger than $\hat{\alpha}_{11}$ $(= 0.174)$. In other words, the panic effect is more important than the reputation effect. Second, $\hat{\alpha}_{12}$ $(= 0.120)$ is much larger than $\hat{\alpha}_{21}$ $(= 0.036)$, which means that existing investors are quite sensitive to large inflows and tend to redeem their shares for fear of performance drop after a large inflow. This highlights the importance of monitoring the purchase counts separately from the redemption counts. By contrast, it seems difficult for the fund to attract new investors after large outflows. Finally, the two eigenvalues of the matrix $(\sum_{i=1}^{\infty} i^{-d})A_1$ are approximately 0.580 and 0.260, respectively, which are both smaller than 1. Thus the joint process seems to be stationary.

Let us now compute the conditional pmf. We focus on horizon 1, since the fund manager usually adjusts the positions on a daily basis. For each trading day t in our observation period, we compute $\Pr(X_{1,t+1} = m, X_{2,t+1} = n|\mathcal{F}_t)$ for all $m, n \le M$, where we set M to be the largest past observation $M = \max_{t,j} X_{j,t} = 17$. Then we follow [32] and compute the conditional mode $(\hat{X}_{1,t+1}, \hat{X}_{2,t+1})$ defined by

$$(\hat{X}_{1,t+1}, \hat{X}_{2,t+1}) = \arg \max_{0 \le m,n \le M} \Pr(X_{1,t+1} = m, X_{2,t+1} = n|\mathcal{F}_t).$$

Fig. 6 displays the evolution of the conditional mode $\hat{X}_{2,t}$ against the corresponding realized value $X_{2,t}$ for all the past dates. For expository purpose we have chosen a window of 50 trading days.

Globally, the mode forecast satisfactorily capture the local tendency of the count process, although the realized paths tends to be more erratic.

Diagnostic check. Let us finally conduct some adequacy checks of the estimated model. We first compute the ACF/CCF of the estimated model, as well as some summary statistics of the empirical Pearson residuals [41]. For the first one, since in the paper we have only derived the ACF/CCF for finite p, we resort to Monte Carlo simulation to obtain their approximations in this BINAR(∞) model. Due to space constraint this figure is provided in Appendix A.7. We can remark that globally, the ACF/CCF of the estimated process are quite similar to their empirical counterparts reported in Fig. 5. Moreover, the Pearson residuals seem to be well uncorrelated across different lags.

6. Conclusion

We have extended the BINAR(1) model to allow for dependent thinning, arbitrary errors, and higher-order dynamics. This family has intuitive interpretations, tractable stationarity properties, and are rather flexible compared to existing models. More importantly, we have derived tractable expressions for the predictive distributions, allowing for likelihood based estimation and non-linear forecasting. The model has been applied to a new application area, i.e., fund liquidity risk.

In the paper we have followed the literature by focusing on bivariate models. Is it possible to extend our model into higher dimensions? From the fund manager's point of view, this can be of interest since different investors-large corporates and bank/insurance companies, say, can have different behavior. If count series, two for each client category, becomes available, then a multivariate analysis allows to study cross sectional dynamic effects between different clients and may improve the quality of forecasts. The answer to this question is (partially) affirmative, since Eq. (2) can be extended to the multivariate case, using the multivariate Bernoulli distribution [8].

However such extensions are not without downsides. First, this extended model is not closed under margins; for example the bivariate margins of the trivariate extension no longer have representation (2), except when the matrix A is diagonal; see [35]. Second, and most importantly, as in many other multivariate models, when the dimension increases, both the number of parameters and the computational burden increase. These issues might be mitigated by introducing constrained specifications (see, e.g., Proposition 2), or by conducting pair-wise analysis [21,35]. These await future research.

Acknowledgments

We thank the Editor-in-Chief, Christian Genest, an Associate Editor as well as two referees for helpful comments. Any errors are ours.

Yang Lu acknowledges financial support from the ACPR chair "Regulation and Systemic Risk". Serge Darolles and Gaëlle Le Fol gratefully acknowledge that this research was conducted within the Research Initiative Quantitative Management Research Initiative (QMI) under the aegis of the Fondation du Risque, a joint initiative by Universite Paris-Dauphine, ENSAE ParisTech, Addstones-GFI and La Francaise Investment Solutions. Darolles and Le Foll also acknowledge financial support from the ANR via the Project MultiRisk (ANR-16-CE26-0015-02).

Appendix A

A.1. Proof of Proposition 4

Under the assumptions of the proposition, we have

$$\ln b^{(1)}(u,v) = \lambda_1(u-1)+\lambda_2(v-1) = (\lambda_1,\lambda_2)(u-1,v-1)^\top,$$

where $\lambda_1 = \mathrm{E}(\epsilon_{1,t})$ and $\lambda_2 = \mathrm{E}(\epsilon_{2,t})$. By Eq. (9) we get

$$\ln b^{(\infty)}(u,v) = \sum_{h=1}^{\infty} \ln b^{(h)}(u,v) = \sum_{h=1}^{\infty}(\lambda_1,\lambda_2)(A^\top)^h(u-1,v-1)^\top = (\lambda_1,\lambda_2)(\mathrm{Id}-A^\top)^{-1}(u-1,v-1)^\top,$$

which is the log pgf of a bivariate independent Poisson distribution, with expectations $(\lambda_{1,\infty},\lambda_{2,\infty})^\top = (\mathrm{Id}-A)^{-1}(\lambda_1,\lambda_2)^\top$.

Let us now check that $(Y_t) = (X_{1,t}+X_{2,t})$ follows a Poisson INAR(1) process. The joint pgf of $(X_{1,t-1}+X_{2,t-1}, X_{1,t}+X_{2,t})$ is

$$\begin{aligned}
&\mathrm{E}(u^{X_{1,t-1}+X_{2,t-1}} v^{X_{1,t}+X_{2,t}}) \\
&\quad= \mathrm{E}\Big[u^{X_{1,t-1}+X_{2,t-1}}\big\{1+(\alpha_{11}+\alpha_{21})(v-1)\big\}^{X_{1,t-1}}\big\{1+(\alpha_{12}+\alpha_{22})(v-1)\big\}^{X_{2,t-1}} e^{(\lambda_1+\lambda_2)(v-1)}\Big] \\
&\quad= \exp\Big[\lambda_{1,\infty}\big\{u\big(1+(\alpha_{11}+\alpha_{21})(v-1)\big)-1\big\}+\lambda_{2,\infty}\big\{u\big(1+(\alpha_{12}+\alpha_{22})(v-1)\big)-1\big\}+(\lambda_1+\lambda_2)(v-1)\Big] \\
&\quad= \exp\Big[\{\lambda_{1,\infty}(\alpha_{11}+\alpha_{21})+\lambda_{2,\infty}(\alpha_{12}+\alpha_{22})\}uv + u\underbrace{\{\lambda_{1,\infty}(1-\alpha_{11}-\alpha_{21})+\lambda_{2,\infty}(1-\alpha_{12}-\alpha_{22})\}}_{=\lambda_1+\lambda_2 \text{ by Eq. (15)}} \\
&\qquad\qquad + (\lambda_1+\lambda_2)(v-1)-\lambda_{1,\infty}-\lambda_{2,\infty}\Big]
\end{aligned} \tag{A.1}$$

A quick calculation shows that for a univariate Poisson INAR(1) process Z_t satisfying $Z_t = \alpha\circ Z_{t-1}+\eta_t$, where $\alpha\circ$ is the univariate binomial thinning operator, and η_t is i.i.d. $\mathcal{P}(\lambda)$ distributed, the joint pgf is

$$\mathrm{E}(u^{Z_t}v^{Z_{t-1}}) = \exp\Big\{\frac{\alpha\lambda}{1-\alpha}uv+\lambda u+\lambda v-\lambda\Big(1+\frac{1}{1-\alpha}\Big)\Big\}. \tag{A.2}$$

By matching Eqs. (A.1) and (A.2), we conclude that $X_{1,t}+X_{2,t}$ follows a Poisson INAR(1) with $\lambda = \lambda_1+\lambda_2$, and $\alpha = 1-(\lambda_1+\lambda_2)/(\lambda_{1,\infty}+\lambda_{2,\infty})$.

A.2. Proof of Proposition 5

The marginal expectation is obtained by taking expectation in Eq. (5). By the covariance decomposition formula, we get

$$\begin{bmatrix} \mathrm{var}(X_{1,t}), & * \\ \mathrm{cov}(X_{1,t}, X_{2,t}), & \mathrm{var}(X_{2,t}) \end{bmatrix} = \begin{bmatrix} \mathrm{var}(\alpha_{11}X_{1,t-1} + \alpha_{12}X_{2,t-1}), & * \\ \mathrm{cov}(\alpha_{11}X_{1,t-1} + \alpha_{12}X_{2,t-1}, \alpha_{21}X_{1,t-1} + \alpha_{22}X_{2,t-1}), & \mathrm{var}(\alpha_{21}X_{1,t-1} + \alpha_{22}X_{2,t-1}) \end{bmatrix}$$
$$+ \mathrm{E}\begin{bmatrix} \mathrm{var}(X_{1,t}|\boldsymbol{X}_{t-1}), & * \\ \mathrm{cov}(X_{1,t}, X_{2,t}|\boldsymbol{X}_{t-1}), & \mathrm{var}(X_{2,t}|\boldsymbol{X}_{t-1}) \end{bmatrix}$$
$$= A\begin{bmatrix} \mathrm{var}(X_{1,t}), & * \\ \mathrm{cov}(X_{1,t}, X_{2,t}), & \mathrm{var}(X_{2,t}) \end{bmatrix} A^\top + \gamma_\epsilon + \mathrm{E}(X_{1,t})\gamma_{12} + \mathrm{E}(X_{2,t})\gamma_{34}.$$

Solving this linear matrix equation yields solution (16). Finally, Eq. (17) is a direct consequence of Eq. (5).

A.3. The first two marginal moments of BINAR(p) model

Under the stationarity condition, the marginal expectation of a BINAR(p) model satisfies

$$\mathrm{E}(\boldsymbol{X}_t) = \sum_{i=1}^{p} A_i \mathrm{E}(\boldsymbol{X}_t) + \mathrm{E}(\boldsymbol{\epsilon}_t) \quad \Leftrightarrow \quad \mathrm{E}(\boldsymbol{X}_t) = (Id - \sum_{i=1}^{p} A_i)^{-1}\mathrm{E}(\boldsymbol{\epsilon}_t).$$

where matrix $Id - \sum_{i=1}^{p} A_i$ is invertible by conditions (19) and (20). The covariance matrix can be obtained using the companion CaR(1) form. More precisely, let us denote by V the $2p \times 2p$ covariance matrix of $(\boldsymbol{Y}_t) = (\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p+1})$, viz.

$$V = \mathrm{E}\Big[\{\boldsymbol{Y}_t - \mathrm{E}(\boldsymbol{Y}_t)\}\{\boldsymbol{Y}_t - \mathrm{E}[\boldsymbol{Y}_t]\}^\top\Big] = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \cdots & \Gamma(p-1) \\ \Gamma^\top(1) & \Gamma(0) & \cdots & \Gamma(p-2) \\ \cdots & \cdots & \cdots & \cdots \\ \Gamma(p-1)^\top & \Gamma(p-2)^\top & \cdots & \Gamma(0) \end{bmatrix},$$

where $\Gamma(h)$ is the auto-covariance function of $(\boldsymbol{X}_t)$ at lag h. Since $\boldsymbol{Y}_t$ has the VAR(1) representation

$$\mathrm{E}(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}) = \underbrace{\begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I_2 & 0 & \cdots & \cdots \\ 0 & I_2 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{=A} \boldsymbol{Y}_{t-1} + \begin{bmatrix} \mathrm{E}(\boldsymbol{\epsilon}_t) \\ 0 \\ \vdots \\ \vdots \end{bmatrix},$$

we have

$$V = AVA^\top + \begin{bmatrix} \gamma_\epsilon + \sum_{i=1}^{p}\Big\{\mathrm{E}(X_{1,t})\gamma_{12,i} + X_{2,t}\gamma_{34,i}\Big\} & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{A.3}$$

where $\gamma_{12,i}$, $\gamma_{34,i}$ are defined in a similar way as in (6). Thus $V = \sum_{h=0}^{\infty}(A^\top)^h V_0 A^h$, where V_0 is the second matrix term on the right-hand side of (A.3).

A.4. Proof of Proposition 7

We first reformulate the condition given in Proposition 7 in terms of the eigenvalues of A.

Lemma 4. *For any matrix $A = (\alpha_{i,j})_{1\le i,j\le 2}$ with nonnegative entries only, the two following sets of conditions are equivalent:*

(i) $\alpha_{11} < 1, \alpha_{22} < 1$, and $(1-\alpha_{11})(1-\alpha_{22}) > \alpha_{12}\alpha_{21}$.
(ii) The eigenvalues of A are smaller than 1 in modulus.

Proof. See [21]. $\square$

Let $\boldsymbol{\epsilon}_t$ be an i.i.d. sequence of innovations and $(Z_{1,t,i,j}, Z_{2,t,i,j})$, $(Z_{3,t,i,j}, Z_{4,t,i,j})$ i.i.d. bivariate Bernoulli variables for any $t \in \mathbb{Z}$, $i, j \in \mathbb{N}$. Then we define the doubly indexed sequences $(\boldsymbol{X}_t^{(n)}) = (X_{1,t}^{(n)}, X_{2,t}^{(n)})_t^\top$ recursively by

$$\forall_t \quad \boldsymbol{X}_t^{(n)} = \begin{cases} \boldsymbol{\epsilon}_t & \text{if } n = 0, \\ \boldsymbol{\epsilon}_t + \sum_{i=1}^{n} \sum_{j=1}^{X_{1,t-i}^{(n-1)}} \begin{bmatrix} Z_{1,t,i,j} \\ Z_{2,t,i,j} \end{bmatrix} + \sum_{i=1}^{n} \sum_{j=1}^{X_{2,t-i}^{(n-1)}} \begin{bmatrix} Z_{3,t,i,j} \\ Z_{4,t,i,j} \end{bmatrix} & \text{if } n > 0, \end{cases}$$

We will show that for each t, the n-indexed sequence $(\boldsymbol{X}_t^{(n)})_n$ converges almost surely to a limit $\boldsymbol{X}_t$, and that the limiting process $\boldsymbol{X}_t$ satisfies Eq. (24) .

A.5.1. Almost sure and $\mathbb{L}^1(\mathbb{P})$ convergence of $\boldsymbol{X}_t^{(n)}$

By induction (with respect to n), it is easily checked that for each t,

$$X_{1,t}^{(n)} \geq X_{1,t}^{(n-1)} \quad \text{and} \quad X_{2,t}^{(n)} \geq X_{2,t}^{(n-1)} \geq 0.$$

Thus for each t, both n-indexed sequences $(X_{1,t}^{(n)})_n$ and $(X_{2,t}^{(n)})_n$ converge almost surely, to limits, say, $X_{1,t}$ and $X_{2,t}$, respectively. Let us now consider the $\mathbb{L}^1(\mathbb{P})$ convergence. We have

$$\mathrm{E}(\boldsymbol{X}_t^{(n)}) = \mathrm{E}(\boldsymbol{\epsilon}_t) + \sum_{i=1}^{n} A_i \mathrm{E}(\boldsymbol{X}_t^{(n-1)}) \leq \mathrm{E}(\boldsymbol{\epsilon}_t) + \sum_{i=1}^{n} A_i \mathrm{E}(\boldsymbol{X}_t^{(n)}) \leq \mathrm{E}(\boldsymbol{\epsilon}_t) + \underbrace{\Big(\sum_{i=1}^{\infty} A_i\Big)}_{=A_\infty} \mathrm{E}(\boldsymbol{X}_t^{(n)}),$$

where the inequalities hold component-wise. Thus we have $(Id - A_\infty)\mathrm{E}(\boldsymbol{X}_t^{(n)}) \leq \mathrm{E}(\boldsymbol{\epsilon}_t)$. Then we remark that the 2×2 matrix $(Id - A_\infty)$ is invertible and all the entries of its inverse are nonnegative. Thus we can multiply both sides by $(Id - A_\infty)^{-1}$ and deduce that

$$\mathrm{E}(\boldsymbol{X}_t^{(n)}) \leq (Id - A_\infty)^{-1} \mathrm{E}(\boldsymbol{\epsilon}_t)$$

is upper bounded when n increases. Hence $(\boldsymbol{X}_t^{(n)})$ converges also in $\mathbb{L}^1(\mathbb{P})$ to $\boldsymbol{X}_t$ for each t by monotonous convergence theorem.

A.5.2. Strict and second-order stationarity of process $(\boldsymbol{X}_t)_t$

By definition, for each n, process $(\boldsymbol{X}_t^{(n)})_t$ is strictly stationary. Thus the limiting process $\boldsymbol{X}_t$ is also strictly stationary. By the $\mathbb{L}^1(\mathbb{P})$ convergence, $\boldsymbol{X}_t$ is mean-stationary. To show the covariance stationarity, let us check that the sequence $(X_{1,t}^{(n)})_n$ is bounded in $\mathbb{L}^2(\mathbb{P})$. We have

$$\begin{aligned} \mathrm{var}(X_{1,t}^{(n)}) &= \mathrm{var}(\epsilon_{1,t}) + \mathrm{var}\Big(\sum_{i=1}^{n} \sum_{j=1}^{X_{1,t-i}^{(n-1)}} Z_{1,t,i,j} + \sum_{i=1}^{n} \sum_{j=1}^{X_{2,t-i}^{(n-1)}} Z_{3,t,i,j}\Big) \\ &\leq \mathrm{var}(\epsilon_{1,t}) + \mathrm{E}\Big\{\Big(\sum_{i=1}^{n} \sum_{j=1}^{X_{1,t-i}^{(n-i)}} Z_{1,t,i,j} + \sum_{i=1}^{n} \sum_{j=1}^{X_{2,t-i}^{(n-i)}} Z_{3,t,i,j}\Big)^2\Big\} \\ &\leq \mathrm{var}(\epsilon_{1,t}) + \mathrm{E}\Big\{\Big(\sum_{i=1}^{n} \sum_{j=1}^{X_{1,t}^{(n)}} Z_{1,t,i,j} + \sum_{i=1}^{n} \sum_{j=1}^{X_{2,t}^{(n)}} Z_{3,t,i,j}\Big)^2\Big\} \\ &\leq C_1 + \mathrm{var}\Big\{\Big(\sum_{i=1}^{\infty} \alpha_{11,i}\Big) X_{1,t}^{(n)} + \Big(\sum_{i=1}^{\infty} \alpha_{12,i}\Big) X_{2,t}^{(n)}\Big\}, \end{aligned}$$

where the constant C_1 is independent of t and n. Similar upper bounds can also be obtained for $\mathrm{var}(X_{1,t}^{(n)})$ and $\mathrm{var}(X_{2,t}^{(n)})$ and in matrix form we have

$$V_n \leq A_\infty V_n A_\infty^\top + C,$$

where V_n is the covariance matrix of $\boldsymbol{X}_t^{(n)}$, and C is a constant matrix. Thus we get

$$V_n \leq \sum_{i=0}^{\infty} A_\infty^i C (A_\infty^\top)^i,$$

which is uniformly bounded. Thus by the dominated convergence theorem, $(\boldsymbol{X}_t^{(n)})_n$ also converges to $\boldsymbol{X}_t$ in $\mathbb{L}^2(\mathbb{P})$.

A.5.3. Conditional distribution of X_t given its past

It remains to check that the above limiting process $\boldsymbol{X}_t$ satisfies the representation (24). It suffices to show that, for a fixed t, the sequence

$$\boldsymbol{r}_t^{(n)} = \boldsymbol{\epsilon}_t + \sum_{i=1}^{n} \sum_{j=1}^{X_{1,t-i}} (Z_{1,t,i,j}, Z_{2,t,i,j})^\top + \sum_{i=1}^{n} \sum_{j=1}^{X_{2,t-i}} (Z_{3,t,i,j}, Z_{4,t,i,j})^\top$$

converges to $\boldsymbol{X}_t$ almost surely, or equivalently, since

$$\boldsymbol{X}_t - \boldsymbol{r}_t^{(n)} = \boldsymbol{X}_t - \boldsymbol{X}_t^{(n)} + \boldsymbol{X}_t^{(n)} - \boldsymbol{r}_t^{(n)},$$

it suffices to show that $\boldsymbol{r}_t^{(n)} - \boldsymbol{X}_t^{(n)}$ converges to zero almost surely. But this sequence is non-decreasing, thus it suffices to find an almost sure convergent subsequence. Its $\mathbb{L}^1(\mathbb{P})$ norm is

$$\mathrm{E}(\boldsymbol{r}_t^{(n)} - \boldsymbol{X}_t^{(n)}) = \sum_{i=1}^{n} A_i \mathrm{E}(\boldsymbol{X}_t) - \sum_{i=1}^{n} A_i \mathrm{E}(\boldsymbol{X}_t^{(n)}) \longrightarrow 0$$

when $n \to \infty$. Thus $(\boldsymbol{r}_t^{(n)} - \boldsymbol{X}_t^{(n)})_n$ converges in $\mathbb{L}^1(\mathbb{P})$ and admits a subsequence that is almost surely convergent. Thus the process $\boldsymbol{X}_t$ satisfies Eq. (24).

A.5.4. The necessary condition of stationarity

Thus under conditions (25), (26), (27), the BINAR(∞) process exists and is stationary. Let us now show that these conditions are also necessary. Taking expectation in (24), we get

$$\mathrm{E}(\boldsymbol{X}_t) = \mathrm{E}(\boldsymbol{\epsilon}_t) + A_\infty \mathrm{E}(\boldsymbol{X}_t), \tag{A.4}$$

thus all the entries of A_∞ are finite, hence inequality (25). Finally, by iteration we have

$$\forall_{n\geq 1} \quad \mathrm{E}(\boldsymbol{X}_t) = (Id + A_\infty + A_\infty^2 + \cdots + A_\infty^n)\mathrm{E}(\boldsymbol{\epsilon}_t) + A_\infty^{n+1}\mathrm{E}(\boldsymbol{X}_t).$$

Thus the largest eigenvalue of A_∞ is smaller than 1 in modulus. Then by Lemma 4 we get conditions (26) and (27).

A.5. Proof of Lemma 2

Under Assumption 1, each $\delta_i(\mathcal{F}_{t-1})$ lies between 0 and 1, thus $F(i|\mathcal{F}_{t-1})$ is nondecreasing, and bounded above by 1. Its lower limit is positive under the stationarity condition since

$$\begin{aligned} F(0|\mathcal{F}_{t-1}) &= \exp\Big\{\sum_{i=1}^{\infty} X_{1,t-i} \ln(1+q_{1,i}-\alpha_{11,i}-\alpha_{21,i}) + \sum_{i=1}^{\infty} X_{2,t-i} \ln(1+q_{2,i}-\alpha_{12,i}-\alpha_{22,i})\Big\} \\ &\geq \exp\Big\{-\sum_{i=1}^{\infty} X_{1,t-i} \ln(1-\alpha_{11,i}-\alpha_{21,i}) - \sum_{i=1}^{\infty} X_{2,t-i} \ln(1-\alpha_{12,i}-\alpha_{22,i})\Big\}. \end{aligned} \tag{A.5}$$

As $\alpha_{11,i}+\alpha_{21,i} \to 0$ and $\alpha_{12,i}+\alpha_{22,i} \to 0$ when $i \to \infty$, for large i, we have

$$-\ln(1-\alpha_{12,i}-\alpha_{22,i}) > -2(\alpha_{12,i}+\alpha_{22,i}), \qquad -\ln(1-\alpha_{12,i}-\alpha_{22,i}) > -2(\alpha_{12,i}+\alpha_{22,i}).$$

Then since $\sum_{i=1}^{\infty}(\alpha_{12,i}+\alpha_{22,i})X_{1,t-i} + \sum_{i=1}^{\infty}(\alpha_{12,i}+\alpha_{22,i})X_{2,t-i}$ is finite, the right-hand side of (A.5) is positive.

A.6. Proof of Lemma 3

The proof is based on the fact that the process $\boldsymbol{X}_t$ has the weak VAR(∞) representation

$$\boldsymbol{X}_t = \sum_{i=1}^{\infty} A_i \boldsymbol{X}_{t-i} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\eta}_t$ is a weak white noise. Then we revert this VAR(∞) representation into the Vector MA(∞) representation $\boldsymbol{X}_t = (Id + \sum_{i=1}^{\infty} B_i L^i)\boldsymbol{\eta}_t$, where B_j are matrices and L is the lag operator. By mimicking the proof of Theorem 2 in [43], we can show that $B_i = O(i^d)D$ for some constant matrix D and the autocovariance function also decays at the same hyperbolic rate i^d.

A.7. Additional figures

See Figs. 7 and 8.

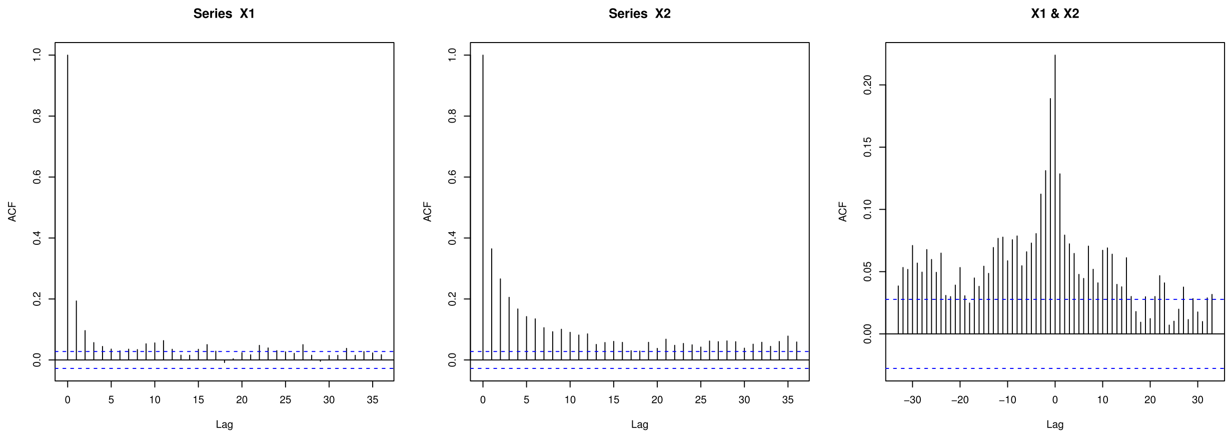

Fig. 7. ACF/CCF of the constrained BINAR(∞) process defined by Eqs. (26) and (31), with parameters given by the MLE estimator.

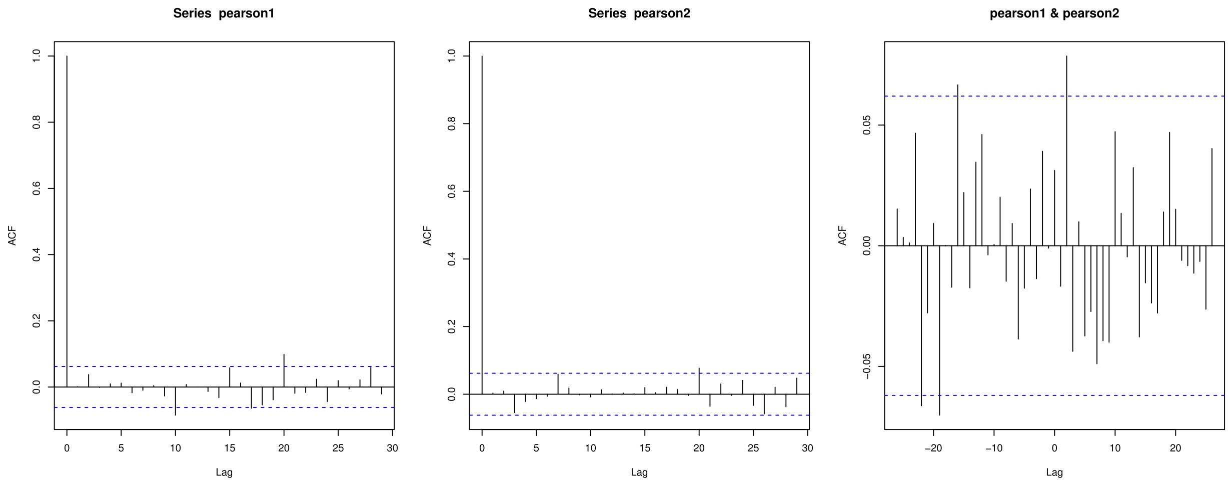

Fig. 8. ACF/CCF of the Pearson's residuals computed using the observed data and the same estimated model as in Fig. 7.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.02.015.

References

[1] Securities and Exchange Commission, Open-End Fund Liquidity Risk Management Programs; Swing Pricing; Re-Opening of Comment Period for Investment Company Reporting Modernization Release, 2015.
[2] K. Bien, I. Nolte, W. Pohlmeier, An inflated multivariate integer count hurdle model: An application to bid and ask quote dynamics, J. Appl. Econometrics 26 (2011) 669–707.
[3] R. Blundell, R. Griffith, J. Van Reenen, Market share, market value and innovation in a panel of British manufacturing firms, Rev. Econ. Stud. 66 (1999) 529–554.
[4] U. Böckenholt, Mixed INAR(1) Poisson regression models: Analyzing heterogeneity and serial dependencies in longitudinal count data, J. Econometrics 89 (1998) 317–338.
[5] M. Boudreault, A. Charpentier, Multivariate Integer-Valued Autoregressive Models Applied to Earthquake Counts, UQAM DP, arXiv preprint arXiv:1112.0929, 2011.
[6] K. Brännäs, J. Hellström, J. Nordström, A new approach to modelling and forecasting monthly guest nights in hotels, Internat. J. Forecast. 18 (2002) 19–30.
[7] R. Bu, B.P. McCabe, Model selection, estimation and forecasting in INAR (p) models: A likelihood-based Markov Chain approach, Internat. J. Forecast. 24 (2008) 151–162.
[8] N.R. Chaganty, H. Joe, Range of correlation matrices for dependent Bernoulli random variables, Biometrika 93 (2006) 197–206.
[9] Y. Cui, F. Zhu, A new bivariate integer-valued garch model allowing for negative cross-correlation, TEST 27 (2018) 428–452.
[10] S. Darolles, Liquidity risk and investor behavior: Issues, data and models, Autorité Marchés Financiers Sci. Advisory Board Review (2018).
[11] S. Darolles, C. Gouriéroux, J. Jasiak, Structural Laplace transform and compound autoregressive models, J. Time Series Anal. 27 (2006) 477–503.
[12] S. Desmettre, J. de Kock, P. Ruckdeschel, F.T. Seifried, Generalized pareto processes and liquidity, Quant. Financ. 18 (2018) 1327–1343.

[13] P. Doukhan, K. Fokianos, B. Støve, D. Tjøstheim, Multivariate Count Autoregression, Université de Cergy-Pontoise DP, arXiv:1704.02097, 2017.
[14] J. Du, Y. Li, The integer-valued autoregressive (INAR(p)) model, J. Time Series Anal. 12 (1991) 129–142.
[15] C.B. Edwards, J. Gurland, A class of distributions applicable to accidents, J. Amer. Statist. Assoc. 56 (1961) 503–517.
[16] C. Genest, J. Nešlehová, A primer on copulas for count data, ASTIN Bull. 37 (2007) 475–515.
[17] C. Genest, J. Nešlehová, M. Ruppert, Comment on "statistical models and methods for dependence in insurance data" by s. haug, c. Klüppelberg, l. peng, J. Korean Statist. Soc. 40 (2011) 141–148.
[18] L. Giraitis, P. Kokoszka, R. Leipus, Stationary ARCH models: Dependence structure and central limit theorem, Econometric Theory 16 (2000) 3–22.
[19] M.B. Gordy, Saddlepoint approximation of CreditRisk$^+$, J. Banking Finance 26 (2002) 1335–1353.
[20] C. Gouriéroux, J. Jasiak, Heterogeneous INAR(1) model with application to car insurance, Insur. Math. Econom. 34 (2004) 177–192.
[21] C. Gouriéroux, Y. Lu, Negative binomial autoregressive process with stochastic intensity, J. Time Ser. Anal. 40 (2019) 225–247.
[22] A. Heinen, E. Rengifo, Multivariate autoregressive modeling of time series count data using copulas, J. Empir. Financ. 14 (2007) 564–583.
[23] H. Joe, Multivariate Models and Multivariate Dependence Concepts, CRC Press, Boca Raton, FL, 1997.
[24] R.C. Jung, A. Tremayne, Coherent forecasting in integer time series models, Internat. J. Forecast. 22 (2006) 223–238.
[25] C. Kemp, H. Papageorgiou, Bivariate hermite distributions, Sankhyā A 44 (1982) 269–280.
[26] M. Kirchner, Hawkes and INAR (∞) processes, Stoch. Process. Appl. 126 (2016) 2494–2525.
[27] A. Latour, The multivariate GINAR (p) process, Adv. Appl. Probab. 29 (1997) 228–248.
[28] H. Liu, Some Models for Time Series of Counts (Ph.D. thesis), Columbia University, New York, 2012.
[29] J. Livsey, R. Lund, S. Kechagias, V. Pipiras, Multivariate integer-valued time series with multivariate flexible autocovariances and their application to major hurricane counts, Ann. Appl. Statist. 12 (2018) 408–431.
[30] Y. Lu, The term structure of predictive distributions is solvable for thinning-based count processes, SSRN ejournal 3095219, 2018.
[31] A.W. Marshall, I. Olkin, A family of bivariate distributions generated by the bivariate Bernoulli distribution, J. Amer. Statist. Assoc. 80 (1985) 332–338.
[32] B.P. McCabe, G.M. Martin, Bayesian predictions of low count time series, Internat. J. Forecast. 21 (2005) 315–330.
[33] B.P. McCabe, G.M. Martin, D. Harris, Efficient probabilistic forecasts for counts, J. R. Stat. Soc. Ser. B 73 (2011) 253–272.
[34] E. McKenzie, Some simple models for discrete variate time series, J. Amer. Water Res. Assoc. 21 (1985) 645–650.
[35] X. Pedeli, D. Karlis, On composite likelihood estimation of a multivariate INAR(1) model, J. Time Ser. Anal. 34 (2013) 206–220.
[36] X. Pedeli, D. Karlis, Some properties of multivariate INAR(1) processes, Comput. Statist. Data Anal. 67 (2013) 213–225.
[37] A.S. Quoreshi, A bivariate integer-valued long-memory model for high-frequency financial count data, Commun. Statist. Theory Methods 46 (2017) 1080–1089.
[38] L. Schmidt, A. Timmermann, R. Wermers, Runs on money market mutual funds, Amer. Econom. Rev. 106 (2016) 2625–2657.
[39] S. Schweer, C. Wichelhaus, Queueing systems of INAR(1) processes with compound Poisson arrivals, Stoch. Models 31 (2015) 618–635.
[40] M.G. Scotto, C.H. Weiß, M.E. Silva, I. Pereira, Bivariate binomial autoregressive models, J. Multivariate Anal. 125 (2014) 233–251.
[41] C.H. Weiß, An Introduction to Discrete-Valued Time Series, Wiley, New York, 2018.
[42] S. Wicksell, Some theorems in the theory of probability, with special reference to their importance in the theory of homograde correlations, Sven. Aktuarieforeningens Tidskr. (1916) 165–213.
[43] P. Zaffaroni, Stationarity and memory of ARCH (∞) models, Econometric Theory 20 (2004) 147–160.